Hierarchical analysis of multimodal images

Guillaume Tochon

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Signal, Image, Parole et Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par
M. Guillaume TOCHON

Thèse dirigée par **Jocelyn CHANUSSOT** et
codirigée par **Mauro DALLA MURA**

préparée au sein du laboratoire **GIPSA-lab**
dans **l'école doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)**

Analyse hiérarchique d'images multimodales

Thèse soutenue publiquement le **01/12/2015**,
devant le jury composé de :

Jean SERRA
Professeur émérite, ESIEE Paris, Président
Sabine SÜSSTRUNK
Professeur, EPFL, Rapporteur
Thierry GÉRAUD
Professeur, EPITA, Rapporteur
Silvia VALERO
Maître de conférences, Université Paul Sabatier, Examinateur
Paolo GAMBA
Associate Professor, Université de Pavia, Examinateur
José BIOUCAS DIAS
Associate Professor, Instituto Superio Técnico, Examinateur
Jocelyn CHANUSSOT
Professeur, Grenoble-INP, Gipsa-Lab, Directeur de thèse
Mauro DALLA MURA
Maître de conférences, Grenoble-INP, Gipsa-Lab, Co-directeur

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE EEATS

Electronique, Electrotechnique, Automatique, Traitement du Signal

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble Alpes
Mention : Signal, Image, Parole, Télécoms

Présentée et soutenue par

Guillaume TOCHON

Analyse hiérarchique d'images multimodales

Thèse dirigée par Jocelyn CHANUSSOT et Mauro DALLA MURA

préparée au laboratoire Grenoble Images Parole Signal Automatique Laboratoire (Gipsa-Lab)

soutenue le 01/12/2015

Jury :

Président :	Jean SERRA	-	ESIEE Paris, France
Rapporteurs :	Sabine SÜSSTRUNK	-	EPFL, Suisse
	Thierry GÉRAUD	-	EPITA, France
Examinateur :	Silvia VALERO	-	Université Paul Sabatier, France
	Paolo GAMBA	-	Université de Pavia, Italie
	José BIOUCAS DIAS	-	Instituto Superio Técnico, Portugal
Directeur :	Jocelyn CHANUSSOT	-	GIPSA-lab, Grenoble-INP, France
Co-directeur :	Mauro DALLA MURA	-	GIPSA-Lab, Grenoble-INP, France

Acknowledgments

I have been picturing myself writing those acknowledgments quite a few times in the past three years, anticipating to this point where, as I thought, this thesis work would be over. And today, here I am, realizing that it is far from being over. Actually, this is just the beginning of something even bigger, the starting point of an exciting journey in the world of signal processing. Being more skillful to manipulate numbers than words, I have also been wondering a lot about my ability to choose the right words to express those acknowledgments. Well, the simplest words are probably the best ones.

The first and foremost person I want to acknowledge is you, Jocelyn. Thank you for having attracted me six years ago into this wonderful world of signal and image processing. I sometimes wonder whether I would be doing today if I had not come and knocked at your door back then, when I was a first year engineering student deep in thought about his future, and even more importantly, if you had not taken the time to guide me at this moment. Ever since, your always right advice have allowed me to grow up and evolve as a researcher. You have been, and still are, an extraordinary source of motivation and inspiration, and it has been a great pleasure working with you as a supervisor. I want to express to you my sincere gratitude for everything you have brought to me, both from a professional and personal point of views, and notably those incredible opportunities at Stanford and UCLA. In short, thank you for everything!

The next one on the list is you, Mauro. It has been really nice revising my French grammar with you, up to this point where you are now probably outperforming me  . More seriously, I could not have hoped for a better pair than Jocelyn and you to guide me through the ups and downs of this three years marathon, and create such a fruitful environment for me to develop and carry out my own ideas. Still, your were always here to reflect with me and rectify whenever it was needed, and I thank you for your guidance and all those scientific (or not) discussions we have had. Probably this is not over yet!

I would also like to thank my Committee members, starting with Sabine Süsstrunk and Thierry Géraud, who have accepted to evaluate my work by carefully reviewing this manuscript, and for their insightful comments. I thank Jean Serra for his kindness for agreeing to chair this thesis Committee and for his valuable remarks and suggestions. Thank you also to Silvia Valero (it was a pleasure meeting you, finally!), José Bioucas Dias and Paolo Gamba for having accepted to be part of this thesis Committee and travel to Grenoble to attend the defense (or in the case of Paolo, having accepted to experiment a novel videoconferencing method).

I have had the chance to meet and collaborate with a lot of outstanding people throughout this thesis, without whom this work would not be what it currently is. In particular, a huge thank you to Miguel Veganzones, my unofficial third supervisor, for those countless discussions about pretty much everything that is presented in this manuscript, ranging from local spectral unmixing to hierarchical representations and braids of partitions, including Hotelling's T-square statistics and lattice theory. No doubt this work has widely benefited

from your wealth of knowledge. Among all the other collaborators, I would like to thank, from Carnegie at Stanford, Greg Asner for the internship I have had the chance to make with him, which triggered this multimodal image analysis problematic and gave birth to this thesis, and Jean-Baptiste for his patience and help during this internship. I still owe you some beers! 😉 From UCLA, I thank Andrea Bertozzi for having made possible my stay over there, as well as her help, alongside with Jérôme's, while I was struggling with the US Immigration department to eventually get this damn right stamp on my visa...

Now moving on to the GIPSA-lab, such a great place to work! I would like to thank my office mates, Carole, Vincent (who left us too early), Julien (although too rarely here) and then replaced by Raph and Miguel (repeat after me: MigUel, not MigOUel 😉) for our numerous coffees breaks, discussions, and me bothering you with my always starting *"Hé, tu sais comment faire..."* most of the time followed by a computer-related question. It was a pleasure sharing this good old D1140 office with you. I would also like to thank what I could summarize as the *team coinche*: Cindy, Céline, Aude, Tim, Arnaud, Pascal (don't worry, I'll have my revenge at you-know-what!), Florian, Manu, Robin, the other Tim (whatever the order), Alexis, Taia, Lucas (my rolling-over-buffets mate 😉), Quentin, and all the other I have forgotten here (sorry about that). Thank you for those countless *coinches*, coffee breaks, and all those ridiculous lunchtime discussions. Thank you also to Delphine and Yang, for supporting me as an internship supervisor. Also, I would like to mention Marie (or Popi, it is up to you) for having supported me almost two years in the same apartment and for our *How I met* evenings. I don't know if you are more keen now to do the dishes (although I can probably guess the answer), but it was nice fighting with you on this point 😉. And finally, thank you Lucia for all your help when it comes to bureaucratic procedures, this thesis would have been way more painful from an administrative point of you without your kindness!

I could not be more grateful to my family, and especially my parents, Isabelle and Philippe for their never weakening love and support for the past 25 years. It is unquestionably thanks to you if I am who I am today. My deepest gratitude also goes to my grand-father, who taught me very early that hard work always pays off. This thesis is dedicated to all of you.

And last but not least, I would like to thank Julie, my dearest Minouche, for her unwavering support from the beginning. Thank you for your patience, your everyday happiness and joy of living have made the last three years way more pleasant than they could have been otherwise. Thank you for standing by my side, no matter what.

Guillaume

Contents

Acronyms

AMSD	Automatic Matched Subspace Detector
ANC	Abundance Non-negativity Constraint
ASC	Abundance Sum-to-one Constraint
BPT	Binary Partition Tree
DFC	Data Fusion Contest
DSM	Digital Surface Model
ECHO	Extraction and Classification of Homogeneous Objects
EIA	Endmember Induction Algorithm
ERGAS	Erreur Relative Globale Adimensionnelle de Synthèse
FCLSU	Fully Constrained Least Squares Unmixing
FNEA	Fractal Net Evolution Approach
GLRT	Generalized Likelihood Ratio Test
GOF	Goodness-of-fit
HSEG	Hierarchical Segmentation
HSI	Hyperspectral Imagery/Image
ICA	Independent Component Analysis
InSAR	Interferometric Synthetic Aperture Radar
IR	Infrared
JHAPL	John Hopkins Applied Physics Laboratory
LiDAR	Light Detection and Ranging
LMM	Linear Mixing Model
LWIR	Long Wave Infrared
MAP	Maximum A Posteriori
MLE	Maximum Likelihood Estimator
MR	Multi Resolution
MRF	Markov Random Field
PC	Principal Component
PCA	Principal Component Analysis
PolSAR	Polarimetric Synthetic Aperture Radar
RMSE	Root Mean Square Error
RNMF	Robust Non-negative Matrix Factorization
SAE	Spectral Angular Error

SAM	Spectral Angle Mapper
SAR	Synthetic Aperture Radar
SID	Spectral Information Divergence
SVM	Support Vector Machine
TES	Temperature Emissivity Separation
ToS	Tree of Shapes
VCA	Vertex Component Analysis
VNIR	Visible and Near Infrared

List of symbols

Operations on sets

$\mathcal{P}(X)$	Set of all subsets of X
$\lvert X\rvert$	Cardinality of set X
$\delta_{\mathsf{SE}}(X)$	Morphological dilation of set X with structuring element SE
$X \Delta Y$	Symmetric difference between sets X and Y
$\mathbb{1}_X$	Indicator function of set X

Operations on vectors and matrices

$\boldsymbol{\mu}_{\mathcal{R}}$	Mean of all pixel values in region $\mathcal{R}$
$\boldsymbol{\mu}^{\star}_{\mathcal{R}}$	Normalized mean of all pixel values in region $\mathcal{R}$
$\lVert\mathbf{x}\rVert_p$	L_p norm of vector $\mathbf{x}$
$\mathbf{x}^T$	Transpose of vector $\mathbf{x}$
$\mathbb{1}_N$	Column vector of 1's of size N
$\lvert\mathbf{X}\rvert$	Determinant of matrix $\mathbf{X}$
$\mathbf{P}^{\perp}_{\mathbf{X}}$	Projection matrix onto $\mathbf{X}^{\perp}$

Image representations

$\mathcal{I} : E \to V$	Image (functional-based representation)
E	Spatial support of image $\mathcal{I}$
$x \in E$	Element of E (pixel)
V	Space of pixel values
$\mathcal{I}(x) \in V$	Pixel value
$\mathbf{X} \in \mathbb{R}^{N \times N_{\mathsf{pix}}}$	Image (matrix-based representation)
N, N_{pix}	Number of bands (of pixels) in image $\mathbf{X}$
$\mathbf{x} \in \mathbb{R}^N$	Pixel value

Hierarchical representations

$\mathcal{T}$	Tree-based representation
H	Hierarchy of partitions
$\mathcal{R} \in H$	Region of the hierarchy of partitions H
π	Cut of the hierarchy of partitions H
$H(\mathcal{R})$	Sub-hierarchy of H rooted at $\mathcal{R}$

$\pi(\mathcal{R})$	Partial partition of $\mathcal{R}$ (cut of $H(\mathcal{R})$)
B	Braid of partitions
H_m	Monitor hierarchy of braid B

Sets

$\mathcal{R}$	Region (subset) of E
$\partial\mathcal{R}$	Boundary of region $\mathcal{R}$
π	Partition of E
Π_E	Set of all partitions of E
$\Pi_E(H)$	Set of all cuts of hierarchy H
$\Pi_E^c(H)$	Constrained set of cuts of hierarchy H
$\Pi_E(H(\mathcal{R}))$	Set of all partial partitions of $\mathcal{R} \in H$
$\Pi_E(B)$	Set of all cuts of braid B
$\Pi_E(H_m)$	Set of all cuts of monitor hierarchy H_m

Energy functions

$\mathcal{E}$	Energy function
$\mathfrak{D}$	Composition law for energy $\mathcal{E}$
$\pi^\star$	Optimal cut of hierarchy H with respect to $\mathcal{E}$
$\pi^\star(\mathcal{R})$	Optimal partial partition of $\mathcal{R} \in H$ with respect to $\mathcal{E}$
$\mathcal{E}^\star(\mathcal{R})$	Energy of the optimal partial partition $\pi^\star(\mathcal{R})$
$\pi_B^\star$	Optimal cut of braid B with respect to energy $\mathcal{E}$
$\mathcal{E}_\lambda$	Energy function parametrized by λ
$\mathcal{E}_\phi$	Goodness-of-fit term of energy $\mathcal{E}_\lambda$
$\mathcal{E}_\rho$	Regularization term of energy $\mathcal{E}_\lambda$
$\pi_\lambda^\star$	Optimal cut of hierarchy H with respect to $\mathcal{E}_\lambda$
$\pi^\diamond$	Global infimum of the energetic lattice

Spectral unmixing

$\mathbf{e}$	Endmember
$\mathbf{E}$	Matrix of endmembers
ϕ	Fractional abundance
$\mathbf{\Phi}$	Matrix of fractional abundances
$\boldsymbol{\eta}$	Additive noise
$\epsilon(\mathbf{x}, \hat{\mathbf{x}})$	Reconstruction error between $\mathbf{x}$ and $\hat{\mathbf{x}}$
Δ	Matrix of endmember distances
W	Matrix of significance credits

Statistics and probabilities

$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ
$\chi^2_{a,b}$	Non-central χ^2 distribution with a degrees of freedom and non-centrality parameter b
$\mathbf{X}^2_{a,b}$	Cumulative distribution function of $\chi^2_{a,b}$
T^2	Hotelling's T-square statistic
$\Sigma, \hat{\Sigma}$	Covariance and sample covariance matrices
$\Lambda(\mathbf{x})$	GLRT statistic for pixel value $\mathbf{x}$
γ_{GLRT}	GLRT threshold
p_D	Probability of detection
p_{FA}	Probability of false alarm

Miscellaneous

$\mathcal{G} = (\mathcal{U}, \mathcal{V})$	Graph with set of vertices $\mathcal{U}$ and set of edges $\mathcal{V}$
$\mathcal{H}_{\mathcal{R}}$	Empirical distribution (histogram) of pixel values in region $\mathcal{R}$
$\omega_{\mathcal{R}}$	Feature associated to region $\mathcal{R}$
$\Omega_{\mathcal{R}}$	Set of features associated to region $\mathcal{R}$
O_t	Object instance at time t
$\hat{O}_t$	Estimate of the object at time t
$C_{t-1,t}$	Change mask between time $t-1$ and time t
$\epsilon(\pi\|\mathcal{I})$	Average goodness-of-fit of partition π with respect to image $\mathcal{I}$

Introduction

The volume of numerical data generated per year in the world has grown from 1.2×10^{21} bytes in 2010 to 2.8×10^{21} bytes in 2012, and is expected to reach around 40×10^{21} bytes in 2020. This exponential increase, due to the quantitative explosion of the amount of recording sensors, is supported by a worldwide numerical storage capacity which has roughly doubled every 40 months since the 1980s [90], placing us nowadays at the heart of the *digital age*. A major consequence of this tremendous numerical sampling of the real world is the phenomenon of *multimodality*. As a matter of fact, there is no longer a unique sensor devoted to the monitoring of a given physical source, but rather a multiplicity of them thanks to the proliferation of cheap information sensing devices with more advanced sensitivity and specificities, each picturing a particular aspect of the source. This multimodality of recorded signals yields a wealth of information, allowing to better represent and comprehend the sensed scene, but raises on the other and several issues regarding its optimal exploitation.

In particular, there is a real need for adapted multimodal processing tools, as they tend to become more and more widespread in signal and image processing and analysis. Despite their great interest for a broad range of applications is recognized, the huge diversity of multimodal signals makes their general representation a real challenge. While their processing techniques have been so far mostly conditioned by the application field in which they occur, there is a growing interest in the design of new methods to handle multimodality in a more generic way [59, 106].

As human beings, we are continuously subject to multimodal signals through our five senses. The human brain has the capacity to naturally filter all incoming signals and retain out only the useful information in order to interpret the surrounding world and take the optimal decisions to interact with it. While computers undoubtedly outperform human brains in terms of computational capacity, machines are not yet even close to human performances when it comes to interpretation and decision making. One of the main challenges of research nowadays is to bridge this gap by imitating the operating process of the human brain.

In the field of image processing and analysis for instance, real efforts have been done in the last decades to emulate the human vision. Numerical imaging sensors are now far more powerful than human eyes in many aspects. For instance, the human eyes are sensitive only to wavelengths in the visible domain of the electromagnetic spectrum (between 380 nm and 780 nm), while imaging sensors can be devised to collect information in other portions of this spectrum (such as X-ray imaging sensors used in the field of medical imagery, or infrared thermography operating in the thermal infrared domain).

Progresses have also been made to emulate the human brain cognitive processes in terms of image interpretation and understanding, and hierarchical image representations are one of those advances. As a matter of fact, when analyzing an image, the human brain naturally decomposes it into a set of semantically consistent regions which can be associated with real world objects. Taking an aerial photography of a city for example, one automatically

recognizes buildings, parks, roads, and so on In natural images, such set of coherent regions very often organizes itself in a hierarchical way: regions are ordered from fine to coarse, where coarse regions comprises the fine ones. In the aerial picture of a city, trees are contained in parks, buildings and roads are enclosed in neighborhoods, which are themselves comprised in the whole city. The definition of region of interests is related to the notion of scale of exploration, being the level of details at which the image is analyzed. As an image can be explored at various levels of details, the choice of a proper scale of exploration is driven by the underlying application. Coming back to the previous example, one would not operate at the same scale of exploration if the goal was to count the number of trees or cars present in the scene, requiring a fine level of details as each region of interest would be made of a few pixels only, or to evaluate the total length of the road network spanning the whole city, and thus at a coarser representations scale.

Hierarchical representations are a way to accommodate for this intrinsic multiscale nature of images, and have become a popular tool for image analysis that can be adapted to a broad range of applications.

Objectives and thesis organization

This thesis is concerned with the study of multimodality and hierarchical representations. As a matter of fact, the main objective of the work developed here is to connect those two notions. We focus in particular on multimodal images, *i.e.*, several images of the same scene but acquired with different characteristics, such as the type of imaging sensor, the acquisition time, the localization around the imaged source, and so on. This thesis extends hierarchical representations to such multimodal images, in order to exploit at best the information brought by the multimodality and improve the classical image processing techniques when applied to real applications.

The definition of a particular multimodality parameterizes the hierarchical representation of the resulting multimodal image, while the application guides the subsequent processing of this hierarchical representation. Therefore, each chapter of this manuscript is articulated around this quadruplet *multimodality/hierarchical representation/ hierarchical processing/application*, as depicted by figure 1. The overall organization is hierarchical, in the sense that chapter 1 plays the role of the root by introducing the tools on which the following leaf chapters 2, 3 and 4 rely on [1].

Chapter 1 The first chapter introduces the cornerstone notion of multimodality in signal

1. While chapter 1 is the root of the manuscript, the remaining chapters 2, 3 and 4 which constitute the three leaves of the hierarchy depicted by figure 1 are globally uncorrelated in the sense that they all investigate a particular multimodality and implement the tools presented in chapter 1. However, some notions are shared between the three leaf chapters (the BPT representation for hyperspectral images between chapters 2 and 3 and the energetic framework between chapters 2 and 4). While chapters 3 and 4 may be switched (and were organized as such mainly following chronological considerations), it is however strongly advised to read chapter 2 beforehand.

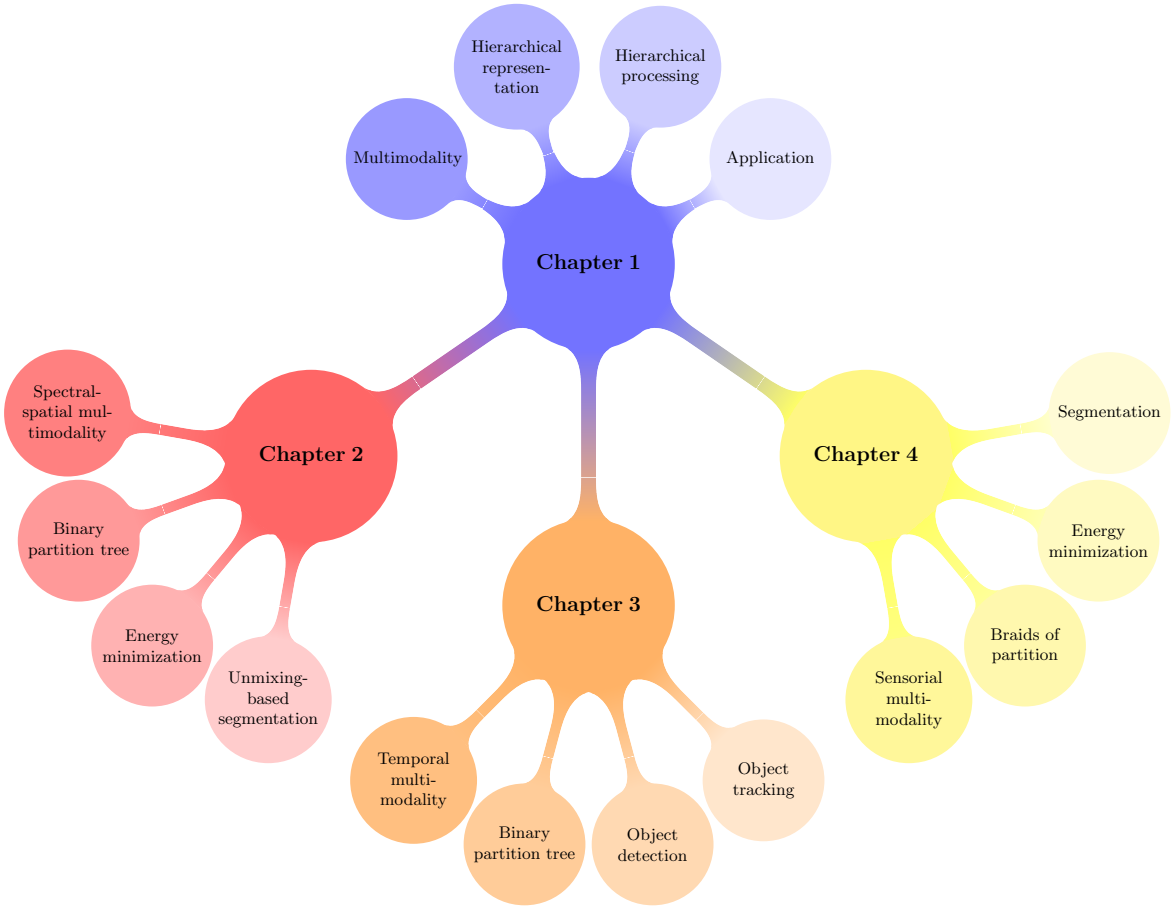


Figure 1: Thesis organization

and image processing, and proposes to model it with a formal definition. A particular focus is notably put in multimodal images frequently encountered in remote sensing. The second keystone notion presented in chapter 1 is the hierarchical representation of images. Theoretical foundations and definitions are first described, and tree-based representation as well as hierarchical representation are then reviewed, the latter being a particular case of the former. The binary partition tree (BPT), considered as the baseline hierarchical representation in the following chapters, is presented more in details. Finally, chapter 1 illustrates the use of hierarchical image representation and analysis in a concrete example being the segmentation of tropical rain forest hyperspectral images. This application also allows to motivate the design of hierarchical multimodal tools and underline their challenges.

Chapter 2 In this second chapter, we focus on the *spectral-spatial* multimodality, naturally provided by hyperspectral images. In particular, the use of spectral and spatial information has already proved to be valuable for spectral unmixing purposes. In opposition to the classical case which perform the unmixing over the global image, we adopt here a local point of view, as we aim at definition a segmentation of the hyperspectral image

that is optimal with respect to the unmixing operation. A first contribution is to propose new strategies to build a BPT representation of hyperspectral images with novel region models and merging criteria, adapted for spectral unmixing purposes. Then, an optimal segmentation is extracted from the BPT structure following some energetic considerations. In particular, attaching some energy function to all segmentations contained in the BPT structure, we search for the one with minimal energy, and therefore maximal adequacy with the intended application. While the energy minimization procedure has already been studied in the literature for certain type of energies [87], a second contribution of this chapter is to extend this procedure to new definitions of energy functions. Based on these results, we finally formulate novel energy functions which aim at producing some segmentation being optimal with respect to the spectral unmixing, reached by combining both the spectral and spatial information contained by the hyperspectral image.

Chapter 3 This third chapter concentrates on the *temporal* multimodality, that is, when the multimodal data features several images acquired at different dates and can be thus assimilated to a video sequence. While the processing of traditional video sequences has been thoroughly investigated in the computer vision literature, we propose to consider the case of hyperspectral video sequences instead. In particular, we focus on the object tracking application, which consists in following the motion of an object of interest as it evolves with time along the sequence. The contribution of this chapter is to propose a novel method for object tracking, tackled as a sequential hierarchical object detection procedure. It first involves the construction of a BPT over each frame of the hyperspectral video sequence and then the retrieval of the tracked object of interest among the nodes of the BPT structure. The proposed object tracking method is tested in a real scenario being the tracking of a chemical gas plume in thermal hyperspectral video sequences.

Chapter 4 The last chapter of this manuscript is devoted to the *sensorial* multimodality, *i.e.*, when several images of a scene are acquired with different sensors. This multisource multimodality is appealing in particular for image segmentation applications, as the information brought by the various modalities of the multimodal images should lead to the design of more accurate regions. However, handling each individual image by a hierarchical representation raises the question on how the fusion of those hierarchies. Based on the recently proposed concept of braids of partitions [101] (being an extension of hierarchies of partitions), we derive a novel methodology for the fusion of hierarchical representations. Using again an energetic framework, the contribution of this chapter is to perform the hierarchical segmentation of multisource images using braids of partitions. The validation of the proposed methodology is conducted using various sensorial multimodal data sets.

Chapter 1

Multimodality and hierarchical representations

Contents

The major role of this first chapter is to introduce the two fundamental (and allegedly unrelated) notions that are at the core of this manuscript, namely *multimodality* and *hierarchical representations*. The former concept is presented in section 1.1, which reviews the different cases of multimodality often occurring in signal and image processing (with a particular focus for the remote sensing field). Section 1.2 defines some classical image representations, acting as a prerequisite for section 1.3 which presents common tree-based and hierarchical image representations. Section 1.4 features an example of hierarchical image representation and analysis in a concrete scenario, namely the hierarchical segmentation of tropical rain forest

hyperspectral images, and illustrates both the need for adapted multimodal tools and the challenges their design represent.

1.1 Multimodality

What is multimodality?

The most basic, linguistics-driven answer to this question, is *something which is composed of several modalities*. Then, what is a modality? According to the Cambridge dictionary [1], a modality is *a particular way of doing or experiencing something*. Its Oxford counterpart [2] proposes an equivalent definition, as being *a particular form of sensory perception*. The notion of modality is therefore linked with the notion of signal received by a sensor, and by extension, it can be intuited that multimodality involves several signals and/or several sensors. Using a signal processing terminology, one can reformulate the definition of multimodality as done in [107]: a multimodal signal is defined as the information about some physical phenomenon, or system of interest, recorded with different types of sensors and/or at different locations and/or at different observation times and/or using different experimental setups... Again appears in this definition the notion of multiplicity of signals and receivers, and the intuition that the description and categorization of multimodal signals is rather broad.

What does not appear in the previous definition however, is the interest of such multimodal signals, not necessarily straightforward at a first sight. While the following 160 pages can be considered as a tentative answer to this point, let us begin with an illustrative example. We, as human being, are undergoing multimodal signals in an everyday basis. We continuously record signals and information through our five senses: sight, hearing, smell, touch and taste. Our brain naturally processes and summarizes all this simultaneously recorded information, to retain only the most important part of it, the one that allows us to interact with our external environment. As these processings are done internally and innately, we are not necessarily aware that they actually allow us to perceive the environment at our best. Discard only one of those five senses, and the picture of our environment becomes only partial.

Let us illustrate this assertion with the simple example of a musical concert, where the two dominant working senses would be the sight and the hearing. Obstructing one of these senses (by inserting earplugs or wearing a blindfold) would yield a different and only partial perception of the concert. This perception could also be altered by the localization in the concert room: a person located right in front of the stage would probably not sense the concert the same way than a person located at the rear of the stands, and one could argue that each person experience is only partial. Similarly, attending the concert in live or watching it on TV (*i.e.*, changing the "experimental setup") would provide only a sided information.

From the above example, it should be clear that a multimodal signal bears more information than each of its individual modalities, also called components. A major challenge of nowadays

1. http://dictionary.cambridge.org/dictionary/english/modality
2. http://www.oxforddictionaries.com/definition/english/modality

computer-based technological era is to make computers, with powerful, well understood, and more importantly, well mastered computational capacities, mimicking the human brain, whose computational capacities are even more powerful in some sense, but far from being understood and mastered. Consequently, multimodal data handling and processing is a very active field of the signal and image processing community, and the challenges to be taken up are numerous [107].

1.1.1 Multimodality in signal and image processing

The practical description of multimodality intuited above proposes to define it as the joint consideration of several signals coming from the same source with different acquisition setups. However, a more formal study multimodality of this phenomenon requires a clear framework holding on a baseline definition, which is the purpose of the current section. Defining a *signal* is the first step toward the definition of a multimodal signal.

Definition 1.1 (signal)
A signal is defined as a function

$$\begin{array}{rccc} \mathcal{I}: & E & \longrightarrow & V \\ & x & \longmapsto & \mathcal{I}(x) \end{array} \tag{1.1}$$

In definition (1.1), the word *signal* is used in its broadest sense, as the acquisition of some physical phenomenon, recorded by some sensor. E will be called the *support* of $\mathcal{I}$, whose *elements* are $x \in E$. V is the space where $\mathcal{I}$ takes its values $\mathcal{I}(x)$.

Definition (1.1) is convenient as it adapts to all kind of signals encountered in the signal and image processing fields. If $\mathcal{I}$ is a temporal signal, recorded with a microphone for instance, then its support is the time axis (thus $E \subseteq \mathbb{R}$), its elements x are the sampling points at which the signal was recorded and its values $\mathcal{I}(x)$ correspond to the numerical data recorded by the microphone (the space of values V is probably a subset of $\mathbb{R}$ as well). If $\mathcal{I}$ is an image, then its support E is the pixel grid $E \subseteq \mathbb{Z}^2$ (one will talk of *spatial* support in this case), its elements x are the pixel sites, and V is the space of pixel values $\mathcal{I}(x)$ being $\mathbb{R}^n$ (or even $\mathbb{C}^n$ for certain types of images) without loss of generality.

Jointly considering several of such defined signals yields the following definition of multimodality:

Definition 1.2 (Multimodal signal)
A multimodal signal is defined as any function

$$\begin{array}{rccc} \mathcal{I}: & E_1 \times \cdots \times E_P & \longrightarrow & V_1 \times \cdots \times V_P \\ & (x_1, \ldots, x_P) & \longmapsto & (\mathcal{I}_1(x_1), \ldots, \mathcal{I}_P(x_P)) \end{array} \tag{1.2}$$

where each $\mathcal{I}_i : E_i \to V_i$ is an individual modality composing the multimodal signal $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_P\}$.

Following definition (1.2), a multimodal signal $\mathcal{I}$ is a set of different individual *modalities* $\mathcal{I}_i$, which may not have either the same supports E_i or spaces of values V_i, but who do correspond to the same recorded phenomenon. This very general definition, considered as standard for the rest of this manuscript, encompasses the majority of multimodal phenomena occurring in signal and image processing. Some examples of such possible multimodalities non-exhaustively include:

- Different sensors simultaneously recording some signals. In such case, each modality $\mathcal{I}_i$ corresponds to the acquisition of a given sensor. Those sensors may be identical but physically placed at different locations around the emitting source, they may record signals of different physical natures from the same spot, or may be a combination of the previous two options.
- One source recorded with a single sensor at different acquisition times or with various acquisition setups. For the former, each $\mathcal{I}_i$ corresponds to a given time t while it is linked with a given setup for the latter.

Any other combination of the previously enumerated examples obviously leads to a signal complying with definition (1.2), which can thus be termed as multimodal. Therefore, due to the high diversity of possible multimodalities, expecting to handle them with some universal processing appears as highly unrealistic. Additionally, their respective application domains may also be varied and not related, making the design of generic multimodal processing tools a very challenging task.

Among the various applications within the signal and image processing fields where multimodality can be encountered, we can notably, and non-exhaustively, list:

Audiovisual processing: Coming back to the concert example given at the beginning of section 1.1, it is evident that one would not fully experience a musical concert if wearing either a blindfold or earplugs. As a matter of fact, vision and audition are going along when one listen to somebody talk [67], and several studies have shown the potential of combining both audio and video signals to achieve better speech recognition [50]. Among the major challenges that must be faced, the frame rate of the video source may differ from the rate at which the audio samples are obtained. In addition, the video source can be viewed in itself as a multimodal data. A short review of existing methods for audiovisual processing can be found in [81]. Note that audiovisual data is a precise case where the multimodalies (*i.e.*, the audio and visual signals) of the multimodal signal do not share the same support E_i of definition (1.2).

Sensor networking: A sensor network corresponds to the dissemination of several connected sensors around the source that has to be monitored. Sensor networking is more and more widespread in several application fields of signal processing, such as underwater acoustics [4], seismology [226], glaciology [152] or smart grid designing [88]. The configuration of a sensor network allows to simultaneously record several signals from the monitored source. This profusion of information induces some redundancy between all recorded signals, allowing in particular to reduce errors in the measurements. On the other hand, the complementarity

(a) (b) (c)

Figure 1.1: Illustrations of multimodalities occuring in (a) seismology (image borrowed from [226]), (b) underwater acoustics (image from [180]) and (c) medical imaging [3].

between each signal should ensure to capture every possible interesting information related to the emitting source. However, one of the main challenges that has to be faced is the huge load of generated data, which is problematic to operate the collaboration between all sensors of the network and the information fusion.

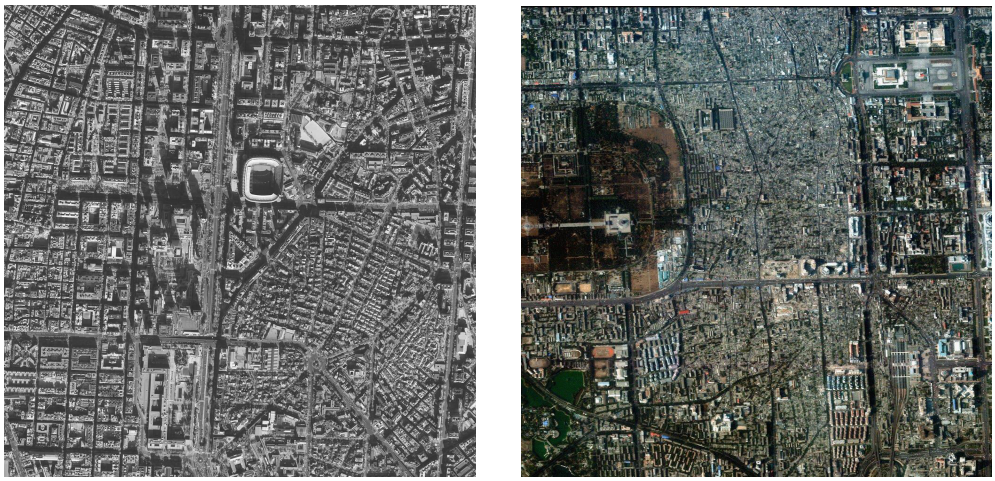
Medical imaging: Medical imaging sensors collect information about organ and tissue anatomy (structural imagery) or their functioning (functional imagery) in order to help practitioners in their diagnoses and assist them during interventions. Due to the highly diverse information that can be acquired, the medical imaging field has seen blossoming a huge quantity of imaging acquisition techniques. Among them, radiography and fluoroscopy are based on X-rays, magnetic resonance imaging and functional magnetic resonance imaging rely on the orientation of the molecules when subject to a strong magnetic field, positron emission tomography detects gamma rays emitted by a chemical tracer introduced in the patient body, and ultrasonography images the echoes made by tissues reflecting pulses of ultrasounds. This wealth of possible multimodalities has led to various studies [37, 187]. One major challenge of medical multimodal images is their co-registration [128] since the various images must be perfectly aligned in order to be fully exploitable by the practitioner.

Illustrations of multimodalities occurring in seismology, underwater acoustics and medical imaging are displayed in figure 1.1.

1.1.2 Multimodality in remote sensing

In its broadest sense, remote sensing consists of the acquisition of information about an object or phenomenon without making physical contact between the sensor and the object of interest. The integration in the past decades of imaging sensors on airborne or spaceborne platforms has made remote sensing a very convenient and well developed technology for Earth observation or, more generally, geoscience applications.

3. http://www.loni.usc.edu/research/projects/OIS/images/multi_modal.jpg



(a) (b)

Figure 1.2: Example of (a) a panchromatic image (QuickBird sensor) and (b) a multispectral image (IKONOS sensor) [4].

1.1.2.1 Imaging sensors in remote sensing

As for the medical imaging field, a important diversity of imaging sensors has been developed for remote sensing applications. While each particular type of sensor is concerned with the measurement of a specific physical quantity emanating from the image scene, sensors used in remote sensing can be categorized in two classes: *passive* sensors, and *active* sensors. The former capture the signal that is emitted by the scene itself (typically, the reflected light) while the latter scan the scene by emitting their own signal and recording the reflected echoes [136].

Panchromatic sensors: Panchromatic images are one-band images, capturing the radiometric information (*i.e.*, the amount of light) that is emitted by the imaged scene in a broad wavelength range (between 450 nm and 900 nm for the IKONOS satellite for instance). Due to their low spectral resolution, they usually produce images at very high spatial resolution (typically less than 1 m per pixel). An example of panchromatic image is displayed in figure 1.2a, featuring the urban area of downtown Madrid (Spain). This image was acquired with the QuickBird satellite, and features a ground resolution of 61 cm. The spectral range covers from 405 nm to 1053 nm.

Multi-spectral sensors: Multi-spectral sensors produce images which are composed of several channels, where the spectral response of each channel is narrower than in the panchromatic case, but remain rather wide (the width being typically around 100 nm). Along with its spatial resolution (which is lower than the one of a panchromatic sensor), a multi-spectral

4. both figures 1.2a and 1.2b are from http://gdsc.nlr.nl/gdsc/en

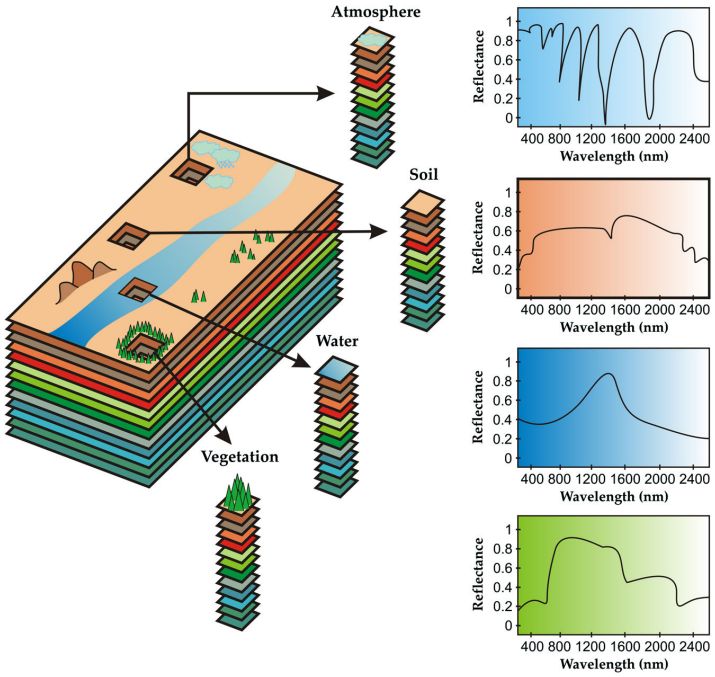
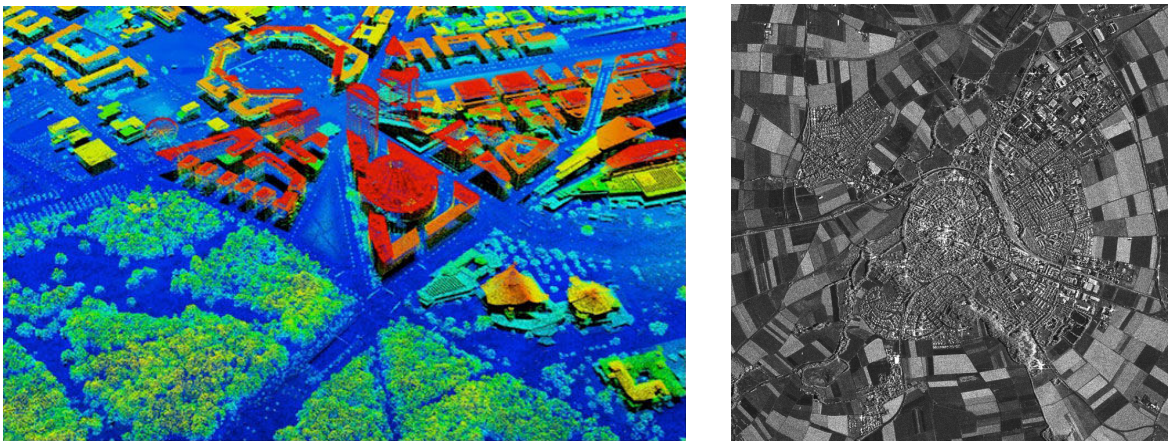


Figure 1.3: Illustration of a hyperspectral image (image borrowed from [23]).

sensor is characterized by its number of spectral channels, their respective width and their location in the electromagnetic spectrum. Multi-spectral sensors mounted on satellites usually contain no more than 10 spectral bands located in the visible domain, typically centered around the blue, green and red domains, and sometimes in the infrared domain, for instance in the near infrared (NIR) and short wave infrared (SWIR) domains. The multi-spectral image displayed in figure 1.2b, acquired over Beijing (China) by the IKONOS satellite, is composed for instance of four spectral bands: blue (450 nm - 520 nm), green (520 nm - 600 nm), red (630 nm - 690 nm) and NIR (760 nm - 900 nm), each of these having a spatial resolution of 4 m. Note that only the red, green and blue bands were used to compose the image of figure 1.2b.

Hyperspectral sensors: Hyperspectral images can be seen as an extension of multi-spectral images as they no longer contain a few spectral channels, but rather up to several hundreds (even thousands in some fields) of them. The spectral channels are in this case centered on a narrow bandwidth, contiguously spaced in the electromagnetic spectrum. Thus, the recorded signal corresponds to a fine sampling of the electromagnetic response of the scene. When this sampling occurs in the visible and NIR domains, this signal can be interpreted as a reflectance function, *i.e.* the function that depicts how the light interacted and was reflected by the imaged scene. When the spectral bands are on the other hand located in the long wave infrared (LWIR) domain, the recorded signal correspond to the emissivity of the scene, *i.e.* the way it has emitted some energy as thermal radiations. Either way, the signal that is recorded for each pixel of the image can be seen as a function of the wavelength, also called



(a) (b)

Figure 1.4: Example of (a) a LiDAR image (color proportional to height) and (b) a SAR X-band intensity image.

spectrum. Moreover, this spectrum, be it reflectance or emissivity, is related to the materials that are physically present in the pixel site. As a matter of fact, two different materials, such as soil and vegetation, do not interact with the light or emit thermal radiations the same way, as depicted by figure 1.3. Hyperspectral images, by their ability to discriminate between materials in the scene, find an always increasing number of applications in the remote sensing field [82, 157]. In order to be able to finely sense the spectral domain and keep a relatively high signal to noise ratio, the spatial resolution of a hyperspectral image needs to be lowered with respect to multi-spectral or panchromatic images: the AVIRIS sensor [85] for instance, produces images composed of 224 spectral bands, of approximately 10 nm width, evenly spaced between 450 nm and 2450 nm, and with a 20 m spatial resolution. The acquisition time could alternatively be extended, but the gain in spatial and spectral resolution would be paid in this case by the presence of a significant motion blur.

LiDAR sensors: As opposed to the panchromatic, multispectral and hyperspectral sensors previously described, which were passive sensors, the light detection and ranging (LiDAR) sensor belongs to the active class. The LiDAR sensor illuminate the scene with a beam of laser and records the time needed for the beam to bounce back from the scene to the sensor. This time lapse is then transformed into a measure of height, called the digital surface model (DSM), as it can be seen in figure 1.4a [5] where hotter colors correspond to higher heights. More advanced LiDAR sensors also record the intensity of the returns, which gives some information about the materials present on the ground, depending on how they reflected the laser pulse (in the same fashion as spectral sensors). Another key characteristic of LiDAR sensors is the wavelength of their emitted pulse. Airborne topographic mapping LiDAR generally use a 1064 nm wavelength laser while bathymetric systems frequently utilize a lower wavelength (such as 532 nm) as it penetrates water with much less attenuation [136]. LiDAR sensors have

5. http://toni88x.bplaced.net/sparse_imgs/lidar2.jpg

found numerous applications in remote sensing, notably in the field of forestry [117] and urban mapping [159].

SAR sensors: Synthetic aperture radar (SAR) is another active sensor commonly used for remote sensing applications. Similar to the LiDAR principle, the SAR sensor illuminates the scene with radio waves and records the echoes reflected by the scene. Depending on how these echoes are processed, one can obtain several information from the radar waves:

Polarimetry: In the case of polarimetric synthetic aperture radar (PolSAR), radio waves are emitted by the sensor with a known polarization (the description of how the electrical component of the emitted electromagnetic wave vibrates in the space). Different materials reflect radar waves with different intensities, but some anisotropic materials also reflect different polarizations with different intensities. By emitting and receiving selective polarizations (for instance, emitting a horizontally polarized wave and receiving it with a vertical polarization), it is then possible to draw a picture of the materials composing the scene.

Interferometry: While PolSAR uses information about wave polarization, the interferometric synthetic aperture radar (InSAR) uses the information contained in the phase of the echoed waves. More precisely, it uses the differential phase of the echoed waves, either from multiple passes along the same trajectory and/or from multiple displaced antennas on a single pass. The processing of this differential phase allows to generate maps of surface deformation or digital elevation models.

The frequency of the emitted radio waves is also of importance when operating a SAR sensor, as different frequencies do not behave the same way when interacting with the ground. For instance, low frequency waves (typically around 0.4 GHz, known as the P-band) are preferred for biomass monitoring and hydrological mapping applications, while higher frequencies (9.6 GHz, corresponding to the X-band) provides the best spatial resolution, thus best suited for surveillance. SAR images find several applications in the remote sensing domain such as land use and land cover classification [161] or change detection [17] for instance. An example of SAR image can be seen in figure 1.4b [6].

1.1.2.2 Multimodality in remote sensing

The multiplicity of sensors used in remote sensing gives rise to numerous occurrences of multimodality, which have found to be useful for many practical application scenarios. Among them, we can notably (and non-exhaustively) list:

The spectral-spatial multimodality: It is one of the most studied multimodality related to hyperspectral imagery. A hyperspectral image is a stack of single-band images acquired at different position of the electromagnetic spectrum. Compared to the classical panchromatic images, hyperspectral images not only contain the spatial information encoded by the pixel

6. http://www.geoville.com/images/TerraSAR-X.jpg

intensities, but also the spectral information through the multiple intensity values of each pixel.

One major improvement brought by the spectral-spatial multimodality concerns the classification of hyperspectral images, which is traditionally conducted in a pixel-wise manner (*i.e.*, each pixel is associated with a class given only its spectral properties). Assuming that spatially close pixels are likely to belong to the same class, spectral-spatial classification methods, such as morphological profiles [72] or watershed segmentation [188] have shown to greatly improve the pixel-wise classification results. The reader is referred to section 2.1.2.1 for a short review of spectral-spatial hyperspectral classification techniques.

Another classical hyperspectral application that benefits from the spectral-spatial multimodality is spectral unmixing. This processing assumes that the spectrum of each pixel of the image can be written as a linear combination of the spectra of some reference pixels (called the *endmembers*) weighted by some coefficients (called the *fractional abundances*) that reflect the contribution of each endmember in the pixel, and aims at estimating the endmembers and associated abundances given a hyperspectral image. Assuming that neighboring pixels should be made of similar materials and in similar proportions, the introduction of some spatial information (regularizing the abundance maps with a Markov random field in [68] and with total variation in [94], or in [135] with some spectral clustering prior to the endmember induction step) within the unmixing process has led to better unmixing results than the classical case where no spatial correlations are taken into account. Similarly, the reader is referred to section 2.1.2.2 for a brief review on spectral-spatial hyperspectral unmixing methods.

The temporal multimodality: In the case of multi-temporal data, the scene is imaged (with the same sensor or not) at different time instances. The comparative analysis of the resulting images allows to detect what has changed in the scene during the lapse of time between two consecutive acquisitions. The most common method to conduct such analysis is to compute and process the image difference, either by direct thresholding or by performing some statistical test [17, 31]. Change detection methods applied to multi-temporal data find several applications in remote sensing, such as the monitoring of vegetation changes [167], or the assessment of natural disasters of environment hazards such as floods [123], tsunamis [26] or wildfires [105].

The multi-angle multimodality Multi-angular images are created when the imaging sensor acquires several images of the scene at different positions (and thus different viewing angles) with respect to the scene. For a sensor mounted on a satellite for instance, images are commonly acquired in the nadir direction (the direction normal to the Earth surface). For multi-angular images, the satellite images a scene at different positions during its pass, resulting in different viewing angles as illustrated by figure 1.5.

Multi-angular images find various applications in remote sensing: in [116, 203] for instance, their are used to estimate the height of buildings in urban environments. As a matter of fact,

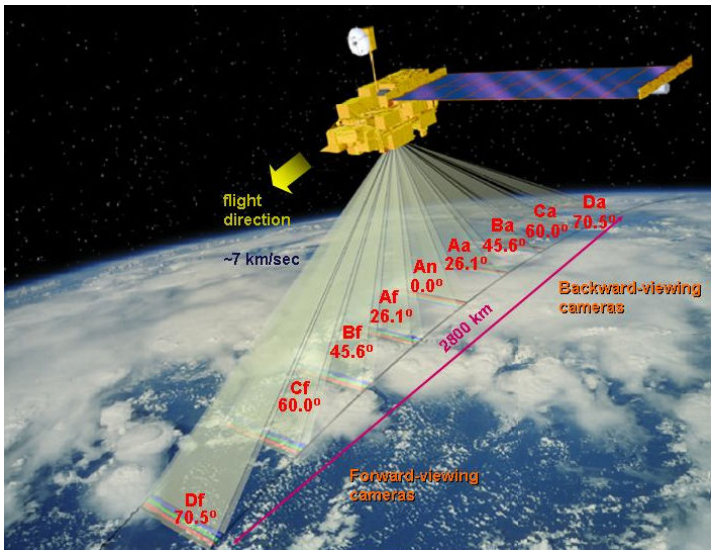


Figure 1.5: Illustration of a multi-angular sensor onboard the Terra satellite [63].

multi-angular images can be viewed and processed as stereoscoptic images, creating a notion of depth for the objects present in the scene. In [49] on the other hand, multi-angular images are used to assess the canopy structure of a boreal forest as the reflectance of various tree species varies differently according to the reflected (scattering) angle. Those variations are therefore captured by the multi-angular images, allowing to discriminate between the tree species composing the forest canopy.

The multisource multimodality Finally, the majority of multimodalities encountered in remote sensing could be classified as multisensor (or multisource) images, which occur when several images of the same scene are acquired with different sensors. One application of the multisensor multimodality is pansharpening, which aims at fusing a high spatial low spectral resolution panchromatic image with a low spatial high spectral resolution multispectral or hyperspectral image in order to create a high spatial high spectral pansharpened image. Several methods, such as component-substitution or multiresolution analysis. Reviews on pansharpening fusion methods and their respective performances can be found in [7, 191]. Of course, the processing of multisensor remote sensing data can pair every type of sources, such as hyperspectral/LiDAR [11, 60] as well as hyperspectral/SAR [48].

1.1.2.3 The Data Fusion Contest

The increasing interest of the remote sensing community toward the processing of multimodal images has led the IEEE Geoscience and Remote Sensing Society to launch in 2006 a data fusion contest (DFC) [7], whose goal is to *evaluate existing methodologies at the re-*

7. http://www.grss-ieee.org/community/technical-committees/data-fusion/

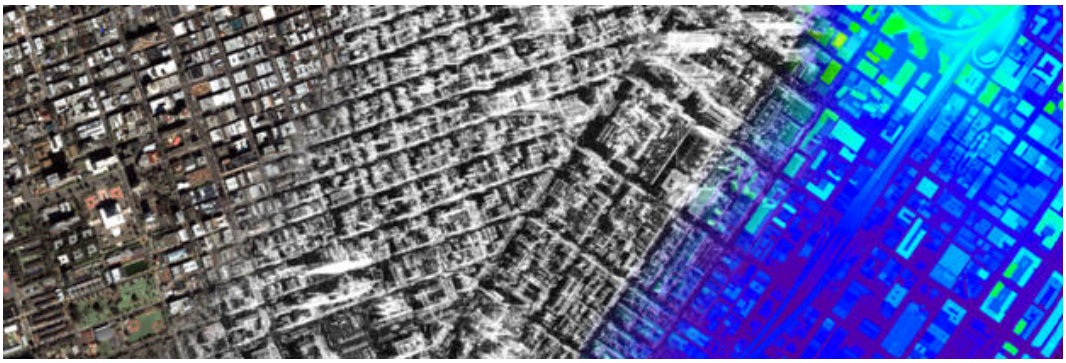


Figure 1.6: Composition of optical (left part), SAR (central part) and LiDAR (right part) images over urban environment [21].

search or operational level to solve remote sensing problems using data from a variety of sensors. This DFC, annually held since 2006, has seen the contenders working on the following multimodalities:

2006: The fusion of multispectral and panchromatic images, also called pansharpening [8].

2007: The fusion of SAR and optical data in an urban mapping framework [151].

2008: The classification of very high spatial resolution hyperspectral data [115].

2009-2010: The analysis of multi-temporal SAR and optical data to perform change detection [123].

2011: The processing of multi-angular panchromatic and multispectral images [150].

2012: The fusion of multi-temporal and multimodal optical, SAR and LiDAR data over some urban environment [21]. A composition of these three modalities is displayed by figure 1.6.

2013: The fusion of hyperspectral and LiDAR data in the framework of hyperspectral classification [61].

2014: The combination of low spatial resolution hyperspectral data acquired in the long wave infrared (LWIR) domain with high resolution optical images [114].

2015: The ongoing DFC 2015 features the processing of very high resolution optical images along with LiDAR data.

The different multimodal data sets and their respective applications proposed in the scope of the DFC illustrates well the numerous instances of multimodalities that can be encountered in the remote sensing field, as well as the need of adapted multimodal tools to process them.

1.1.3 General fusion techniques

The processing of multimodal data necessarily involves, at some time, the pooling of the various features proper to each modality in order to derive some fused features characterizing the multimodal data as a whole. This pooling is classically called the *data fusion step*. Of

course, it depends on the nature of the multimodality being handled, but also on the underlying goal, and there exists no generic method that can be applied regardless of the context. In fact, it would not be exaggerated to say that there exists a specific data fusion step *per* multimodality *per* application. Under this consideration, Wald [223] defined data fusion as a *multilevel, multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from several sources.* A classical data processing chain is composed of three main steps: raw data handling, feature extraction and decision operation. Data fusion can thus take place at three levels of the processing chain [106]:

- Fusion at the raw data level. It is the combination of the raw data from multiple sources into a single "fused" source, which is expected to be more informative than the input sources on their own. A typical example of raw data level fusion is pansharpening, which aims to produce a high spatial high spectral image from a high spatial low spectral and a low spatial high spectral ones.
- Fusion at the feature level. In that case, features of interest (for instance regions, textures, edge maps, and so on) are extracted independently on each source, and are combined to produce some unified feature map that is further used as an input for a single decision step.
- Fusion at the decision level. In such event, features have been extracted and processed on each modality to yield several decision outputs. These decisions are then combined, through majority voting, statistical or fuzzy methods for instance, to produce a final fused decision.

The strategy to adopt, which can be a combination of the previous three fusion techniques, depends of course in practice on the application goal and the type of multimodality to handle.

1.1.4 Conclusion and challenges related to multimodality

As a summary of this first section devoted to multimodality, a *multimodal signal* has been defined as the joint composition of multiple acquisitions of a physical source of interest. Each acquisition procedure, resulting in a given *modality*, differs from one way or another from the other acquisitions, where this difference may be related to the nature of the used sensor, to the position or configuration setup of the sensor, to the date of the acquisition, and so on.

The information carried by each modality is therefore bound to the nature of its acquisition. Several modalities may contain some redundant information: two hyperspectral images acquired at different dates both feature the spectral properties of the materials composing the scene for instance. On the other hand, some other types of information may be explicitly expressed by a single modality of the multimodal data. Therefore, compared to their classical "unimodal" counterparts, multimodal signals are more accurate and more complete representations of the acquired source since they depict multiple facets of it.

While they allow to better describe the recorded source, the processing of multimodal signals is a major issue to their utilization. As a matter of fact, the multimodality phenomenon occurs in various fields of signal and image processing under different natures, thus making the design of generic and portable processing algorithms, even if desirable, highly challenging.

Figure 1.7: Illustration of the pixel grid.

1.2 Image representations and general notations

In the first section of this chapter, the word signal has been used at its broadest sense, as defined by the IEEE Signal Processing Society [8]: *signal refers to any abstract, symbolic, or physical manifestation of information with examples that include: audio, music, speech, language, text, image, graphics, video, multimedia, sensor, communication, geophysical, sonar, radar, biological, chemical, molecular, genomic, medical, data, or sequences of symbols, attributes, or numerical quantities.*

In the following, we focus in particular on images. Multimodal images remain of course concerned by all properties and specificities of multimodal signals defined and discussed in section 1.1. Like any signal, images accept several representations, and each one of them can be processed according to specific mathematical tools. Images are composed of pixels, their smallest structuring element (pixel being a contraction of *pic*ture *el*ement). When an image is digitally stored in a computer, it is represented as a grid map (the *pixel grid*) where one or several values is associated to each cell of the grid, the *pixel values*, as illustrated by figure 1.7. Most of the image representations rely on this grid pattern.

1.2.1 Image as a graph

In the *graph*-based representation, an image $\mathcal{I}$ is depicted as a graph $\mathcal{G} = (\mathcal{V},\mathcal{U})$ where $\mathcal{V}$ is a set of vertices and $\mathcal{U}$ is a set of edges. In such case, each vertex corresponds to a pixel of the pixel grid, as displayed by figure 1.8. The edges, on the other hand, allows to define some neighboring relationships between pixels: two pixels x_i and x_j are neighbors if and only if there is an edge u_{ij} connecting their respective vertices v_i and v_j in the graph $\mathcal{G}$. In particular, the two most common neighboring systems used in image processing are the so-called 4-neighbors and 8-neighbors systems, as represented by figure 1.9. Traditionally, the graph is defined as undirected, meaning that the edges connecting vertices have no directions (if v_i is connected to v_j, then the converse is also true). Pixel values are also stored as attributes for the vertices. One then refers to the graph $\mathcal{G}$ as an undirected vertex-valued graph. Using graph-based

8. http://www.signalprocessingsociety.org/about/scope-mission/

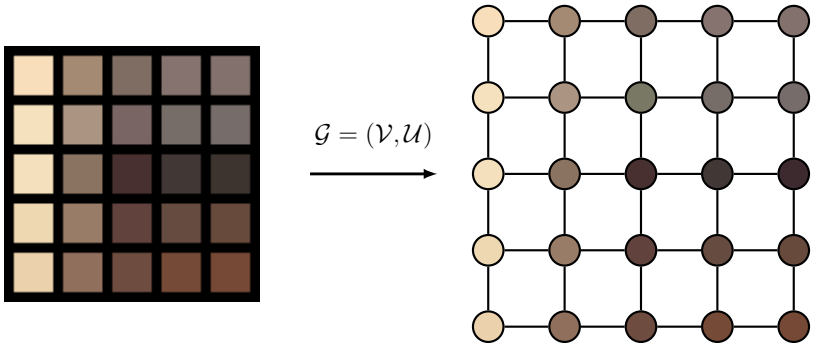
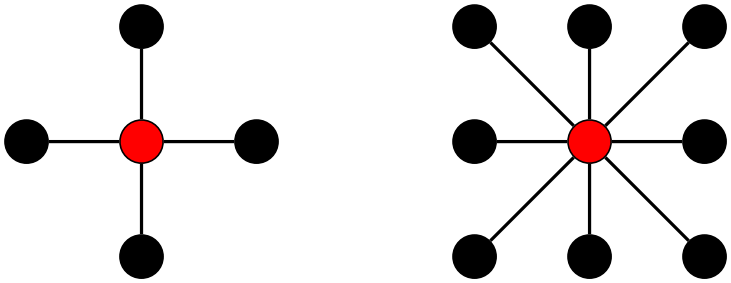


Figure 1.8: Graph-based image representation.



(a) (b)

Figure 1.9: Illustration of (a) 4-neighbors and (b) 8-neighbors systems, where the red central pixel is connected.

image representation notably allows to use tools provided by the graph theory framework, whose most famous examples are probably graph cuts [27] and spectral clustering [222].

1.2.2 Image as a functional

Another possible representation to define an image is as a functional. In such case, we get back to the definition (1.1) of a signal, except that the support spaces can now be more precisely defined. Therefore, in the *functional*-based representation, an image $\mathcal{I}$ is defined as a function

$$\begin{array}{rccc} \mathcal{I}: & E \subseteq \mathbb{Z}^2 & \to & V \\ & x & \mapsto & \mathcal{I}(x) \end{array} \tag{1.3}$$

where E, the *spatial support* of $\mathcal{I}$, is defined as a subset of $\mathbb{Z} \times \mathbb{Z}$ to represent the pixel grid, and the space of values V depends on the used sensor. For a grayscale image (such as a LiDAR image for instance), each pixel value $\mathcal{I}(x)$ is a scalar, thus $V \subseteq \mathbb{R}$. For multi-band images (such as traditional color images, multi-spectral or hyperspectral images), to each pixel is associated a $N-$dimensional vector, where N is the number of channels, and $V \subseteq \mathbb{R}^N$. Note that, in the case of multi-channel images, the pixel vector $\mathcal{I}(x) \in \mathbb{R}^N$ will be denoted by a

bold symbol $\mathbf{x}$ if there is no ambiguity (as it will be the case in chapter 3 notably). In the general case (*i.e.*, for natural images), there is no analytical expression for $\mathcal{I}$ as a function, but properties such as smoothness (at least piece-wise derivability) are commonly assumed. Famous examples of image related quantities relying on the functional-based representation of an image notably include the total variation (TV) [170]

$$TV(\mathcal{I}) = \int_E \|\nabla \mathcal{I}(x)\| dx \ , \tag{1.4}$$

as well as the Mumford-Shah energy functional [142] defined by the further equation (2.8).

1.2.3 Image as a matrix

Another possible image representation is the *matrix*-based formulation. In this case, each pixel value $\mathbf{x} \in \mathbb{R}^N$ of pixel x is considered as a N-dimensional vector, and the whole image $\mathbf{X}$ is viewed as a collection of $N_{\text{pix}} = |E|$ vectors

$$\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_{N_{\text{pix}}}\right] \in \mathbb{R}^{N \times N_{\text{pix}}} \tag{1.5}$$

yielding a matrix constituted of N rows (the dimensionality of the data) and N_{pix} columns (the number of pixels/samples). While the spatial organization of pixels is lost in this matrix-based representation, it makes possible to use classical linear algebra operations such as eigen-decomposition (where a classical example is the principal component analysis, which is applied to the matrix $\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{N \times N}$) or matrix factorization (an example of such matrix factorization, within the framework of hyperspectral unmixing, will be given in section 2.2).

1.2.4 Image as a random vector field

Finally, the *statistical*-based representation considers that each pixel value $\mathbf{x} \in \mathbb{R}^N$ is no longer deterministic, but a particular realization of a more general random variable $\mathbb{X}$ with probability distribution function $p_{\mathbb{X}}$ instead. While simple notions like the mean or the histogram of an image are directly related to this statistical representation, this framework allows for the use of more sophisticated tools. Classical image processing operation, such as classification (with support vector machines [54] for instance), clustering (such as mean shift clustering [52]), object detection and anomaly/target detection (classically involving hypothesis testing [131]) notably exploit the nature and the properties of the probability distribution $p_{\mathbb{X}}$ of the pixel values.

1.3 Hierarchical representations of images

This section is devoted to hierarchical representation of images. Motivated by the intrinsic multi-scale nature of images (section 1.3.1), such representations, also termed *tree-based representations*, have received much attention in image analysis and especially in the field

of mathematical morphology. Sections 1.3.2 and 1.3.3 introduces tree-based representations through this prism of mathematical morphology, while section 1.3.4 focuses deeper into a particular instance of hierarchical representation being the binary partition tree.

The functional-based representation of images is considered in the following, as previously described in section 1.2.2: an image is represented as a function $\mathcal{I} : E \subseteq \mathbb{Z}^2 \to V$.

1.3.1 On the necessity of hierarchical representations

Hierarchical representations are an important and widely used tool in the field of image processing and analysis. Their usefulness come from the intrinsic property of nearly all images to lend themselves well to this type of representations. To understand the reason, we shall take a quick look at the cognitive processes happening in the human brain when examining an image.

As said in the previous section 1.2, an image is digitally stored as a pixel grid, where to each box in the grid is associated a set of values, corresponding to the pixel values. For the computer, there is no relation or connection whatsoever between the values of nearby pixels in the grid. On the other hand, when staring at an image, the human brain does a little bit more than the low-level processing which consists of receipting the electrical signal sent by the optic nerves. It *analyzes* the image and naturally decomposes it into groups of neighboring pixels such that their shape, color or textural attributes have some semantic meaning. Each group can then be linked with a word, and it becomes possible to identify the scene based on the objects from the real world that have been recognized [92].

As an example, consider the image displayed by figure 1.10. For every one who has ever come to Grenoble, it is clear that this picture depicts a nice landscape of this city. The underlying process operated by the brain to recognize the scene is first to divide up the image into regions that are coherent enough to be assigned some semantic meaning (such as "bridge", "river", "building", "mountain", and so on), and then to identify those regions (for instance, the river is recognized as the Isère, the mountains as the Belledonne massif, and so on).

While the process of recognition and identification of regions of interest within an image is natural for the human brain, it is on the other hand one of the most challenging task to mimic in the field of computer vision. Indeed, regions of interest can be defined of various sizes, which is related to the notion of scale of analysis (often referred to as level of details). In the image processing field, images are not just manipulated for fun [9], but because of some underlying application. It is this particular application that dictates the scale at which the image, and thus its regions of interest, should be analyzed. As a simple illustrative example, let us take a second look to the image displayed by figure 1.10. Counting the number of windows or chimneys that appear in this image requires a fine analysis of the scene, and thus a high level of details, since the regions of interest (in other words, the windows or the chimneys) are small, close to the pixel level. On the other hand, enumerating the buildings or separating

9. Although there is some part of fun in it :-)

Figure 1.10: Difference between computer and brain image analyses. For the former, this color image is a pixel grid in which each cell contains three values, the red, green and blue components. For the latter, this is unmistakably a picture of the beautiful city of Grenoble. The brain adds a high-level interpretation process to give some semantic meaning to regions of pixels.

the mountain from the rest of the scene are applications that require a coarser level of analysis because the regions to analyze are significantly larger.

This intrinsic multiscale nature has for consequences that regions of interest are organized in a nested way from fine to coarse scales. Taking back the example of figure 1.10, windows and chimneys are all included in buildings, and all buildings together define the city. Similarly, looking back to figure 1.2 (the panchromatic and multispectral views of a city), one can be interested in extracting buildings individually or neighborhoods, the latter containing the former and thus being of coarser scale, depending on the objective. As the scale of analysis of a single image is bound to the underlying application, it could be useful to decompose an image into all its potential scales of interest regardless of the application, and then browse this collection of scales to choose a proper one, rather than guessing *a priori* for each application what would be the best scale of analysis to operate on. Thus, hierarchical representations appear as a well-suited tool to account for this multi-scale image decomposition. An example of such hierarchical representation is depicted bu figure 1.11.

1.3.2 The lattice of partitions

Working with hierarchical decomposition, and hierarchies of partitions in particular, requires the introduction of mathematical background notions. More specifically and as their name suggests, hierarchies of partitions are composed of partitions. Manipulating partitions is

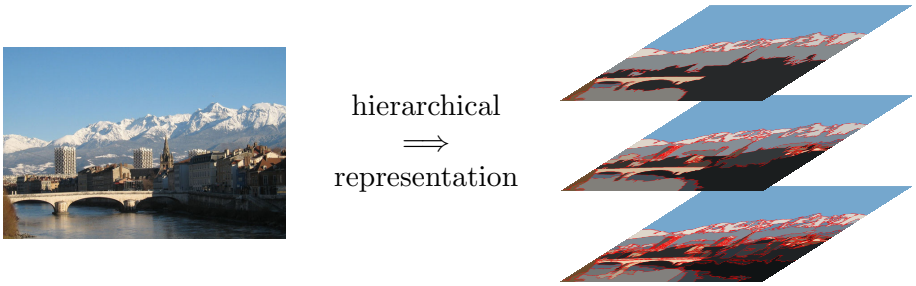


Figure 1.11: Example of decomposition of an image into several scales of interest, represented in a hierarchical structure.

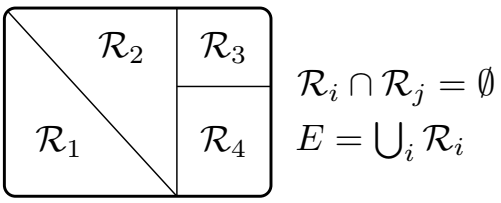


Figure 1.12: Illustration of a partition $\pi = \{\mathcal{R}_i\}$ of a set E.

slightly more complicated than manipulating numbers, and it can be valuable to be familiar with the lattice structure of the space of all partitions.

Definition 1.3 (Partition)
Let E be some set. A partition of E, denoted π, is a family $\{\mathcal{R}_i \subseteq E\}$ of subsets of E such that $\mathcal{R}_i \cap \mathcal{R}_{j\neq i} = \emptyset$ and $\bigcup_i \mathcal{R}_i = E$.

Each subset $\mathcal{R}_i$ is called a *region* (or class) of E. A partition π of E is therefore a division of E into non-overlapping regions which entirely cover E, as illustrates the figure 1.12. The set of all partitions of E is denoted Π_E.

One question that quickly arises when working with partitions is how to compare them. When manipulating numbers, it is natural to use the classical "less than or equal" relation $\leq$. But how does this relation transpose to partitions? Let us first recall the definition of a partial order:

Definition 1.4 (Partial order)
A (non strict) partial order relation on E is a binary operation, denoted $\leq$ which satisfies for any x, y and z in E:

- ***reflexivity:*** $x \leq x$;
- ***transitivity:*** $x \leq y$ *and* $y \leq z \Rightarrow x \leq z$;
- ***antisymmetry:*** $x \leq y$ *and* $y \leq x \Rightarrow x = y$.

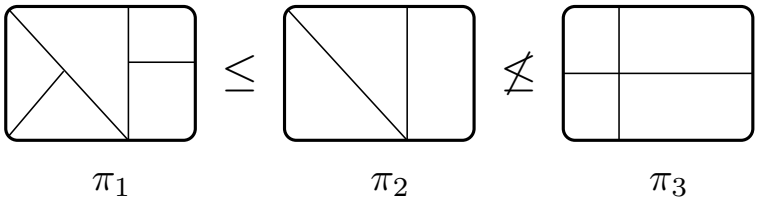


Figure 1.13: Illustration of the refinement ordering: $\pi_1 \leq \pi_2$, but $\pi_1 \not\leq \pi_3$.

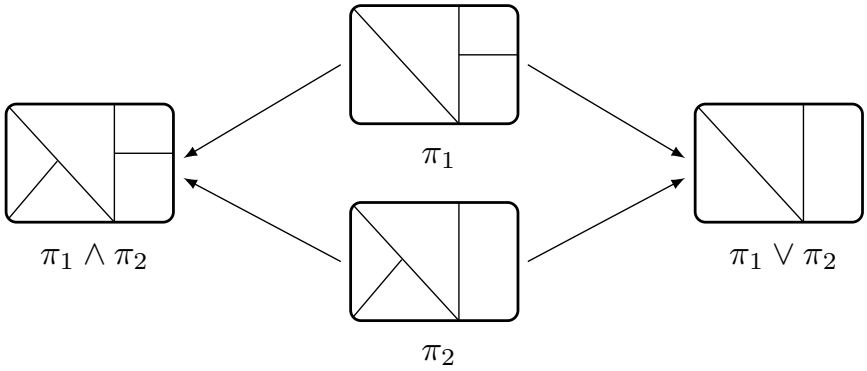


Figure 1.14: Illustration of the refinement infimum (left) and supremum (right) of two partions.

Based on this definition, it is possible to define a partial order on Π_E that reflects the refinement of two partitions:

Definition 1.5 (Refinement ordering)
For any two $\pi_i, \pi_j \in \Pi_E$, one says that π_i refines (or is a refinement of) π_j, and one writes $\pi_i \leq \pi_j$, whenever for each $\mathcal{R}_i \in \pi_i$, there exists $\mathcal{R}_j \in \pi_j$ such that $\mathcal{R}_i \subseteq \mathcal{R}_j$.

In other words, π_i is a refinement of π_j if every individual region $\mathcal{R}_j \in \pi_j$ can be fragmented into one or several regions $\mathcal{R}_i \in \pi_i$, as illustrated by figure 1.13. Informally, π_i is a refinement of π_j if π_i "contains" all the boundaries of π_j. It is for instance the case of π_1 and π_2 in figure 1.13. However, the refinement ordering $\leq$ is only a partial order and not every two partitions are comparable. This is the case in particular for π_1 and π_3 displayed by figure 1.13.

Nevertheless, Π_E equipped with the refinement ordering $\leq$ has a lattice structure, meaning that, even though they are not comparable by refinement, any two partitions π_i and π_j of Π_E always admit a greatest lower bound (called the refinement infimum) $\pi_i \wedge \pi_j$ and a least upper bound (called the refinement supremum) $\pi_i \vee \pi_j$. The former is the largest partition of Π_E that refines both π_i and π_j at the same time, and it is obtained by taking the intersection of all the regions of π_i and π_j. The latter is the smallest partition of Π_E which is refined by both π_i and π_j, and is obtained by retaining only the closed boundaries that are in common

between π_i and π_j. An example of the refinement supremum and infimum of two partitions can be seen in figure 1.14. Notice that the refinement supremum of any two partitions that do not share any closed boundary is the whole support space E. That is the case for instance for partitions π_2 and π_3 of figure 1.13. Conversely, if $\pi_i \leq \pi_j$, then $\pi_i \vee \pi_j = \pi_j$.

Finally, we introduce the notion of sub-lattice, as it will later be useful (in chapter 4 notably):

Definition 1.6 (Sub-lattice)
Let $(E, \leq)$ be a lattice and $E' \subseteq E$. $(E', \leq)$ is a sub-lattice of E if for every two elements x' and y' of E', then $x' \wedge y'$ and $x' \vee y'$ are also in E'.

Put differently, a sub-lattice of E is a subset E' of E such that the supremum and infimum of any two elements of E' are also elements of E'.

1.3.3 Hierarchical representations

1.3.3.1 Tree-based image representation

As it was developed in the previous section 1.3.1, images accommodate well with hierarchical representations since regions of interest within an image are very often either disjoint or nested within each other. To support this observation, tree-based representation have been proposed.

Definition 1.7 (Tree-based representation)
A tree-based representation $\mathcal{T}$ of E is a collection of regions $\mathcal{T} = \{\mathcal{R} \subseteq E\}$ such that:

$$\begin{aligned} &- \emptyset \notin \mathcal{T} \\ &- E \in \mathcal{T} \\ &- \forall \mathcal{R}_i, \mathcal{R}_j \in \mathcal{T}, \mathcal{R}_i \cap \mathcal{R}_j \in \{\emptyset, \mathcal{R}_i, \mathcal{R}_j\} \end{aligned} \tag{1.6}$$

In other words, a tree-based representation of E is a decomposition of E into regions that are either disjoint, or nested. A tree-based representation $\mathcal{T}$ can be represented as a graph $\mathcal{G}_\mathcal{T} = (\mathcal{N}_\mathcal{T}, \mathcal{U}_\mathcal{T})$ where each vertex (also called node) $\mathcal{N}_\mathcal{R} \in \mathcal{N}_\mathcal{T}$ is associated with a region $\mathcal{R} \in \mathcal{T}$ and each edge $u_{i,j} \in \mathcal{U}_\mathcal{T}$ means that either $\mathcal{R}_i \subseteq \mathcal{R}_j$ or $\mathcal{R}_j \subseteq \mathcal{R}_i$. Put differently, the graph representation of $\mathcal{T}$ is the Haase diagram of $\{\mathcal{R} \in \mathcal{T}\}$ ordered by inclusion. In order to simplify the notations, we will denote by $\mathcal{R}$ both the regions of the tree-based representation and the vertices of the associated graph. Based on these tree/graph considerations, we can introduce the following terminology, illustrated by figure 1.15.

Definition 1.8 (Tree terminology)
Let $\mathcal{R} \in \mathcal{T}$. Are defined:

- *The **children** of $\mathcal{R}$ correpond to all regions $\mathcal{R}' \in \mathcal{T}$ that are directly connected to $\mathcal{R}$ in the graph representation of $\mathcal{T}$ and such that $\mathcal{R}' \subseteq \mathcal{R}$. The set of children of $\mathcal{R}$ is denoted $\mathsf{C}(\mathcal{R})$.*

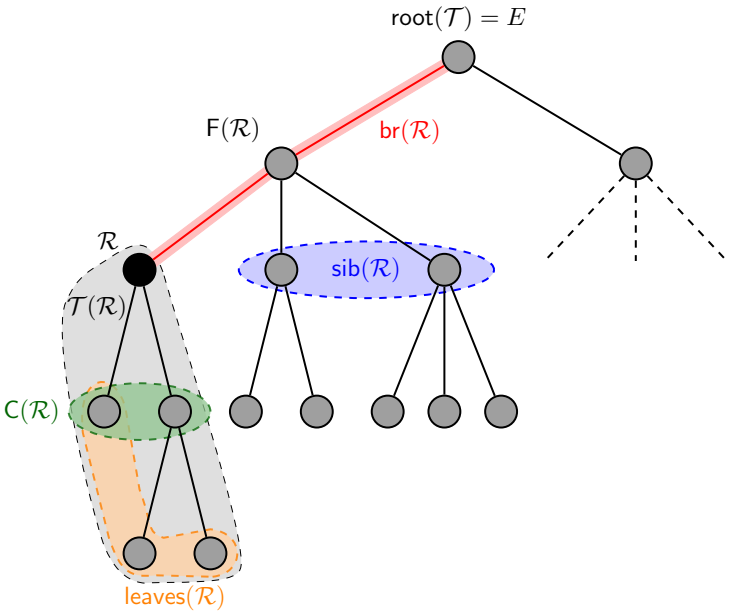


Figure 1.15: Tree-based representation terminology.

- *If $\mathcal{R}$ has no children, i.e., $|\mathsf{C}(\mathcal{R})| = 0$, then $\mathcal{R}$ is called a **leaf** of $\mathcal{T}$. $\mathsf{leaves}(\mathcal{R})$ is the set of leaves of $\mathcal{T}$ that are included in $\mathcal{R}$.*
- *The **father** of $\mathcal{R}$ is, on the other way around, the node $\mathsf{F}(\mathcal{R})$ to which $\mathcal{R}$ is connected and such that $\mathcal{R} \subseteq \mathsf{F}(\mathcal{R})$. In a tree-based representation, each region has exactly one father, except for the root of $\mathcal{T}$, $\mathsf{root}(\mathcal{T})$, which has none.*
- *The **sibling** of $\mathcal{R}$ is the set of regions $\mathsf{Sib}(\mathcal{R})$ that have the same father as $\mathcal{R}$, i.e., $\mathcal{R}' \in \mathsf{Sib}(\mathcal{R}) \Leftrightarrow \mathsf{F}(\mathcal{R}') = \mathsf{F}(\mathcal{R})$.*
- *The **branch** of $\mathcal{R}$, denoted $\mathsf{br}(\mathcal{R})$ is the set of regions $\{\mathcal{R}, \mathsf{F}(\mathcal{R}), \mathsf{F}(\mathsf{F}(\mathcal{R})), \ldots, \mathsf{root}(\mathcal{T})\}$. Elements of $\mathsf{br}(\mathcal{R}) \backslash \{\mathcal{R}\}$ are called **ancestors** of $\mathcal{R}$.*
- *The **height** of $\mathcal{R}$, $\mathsf{h}(\mathcal{R})$, is number of elements in $\mathsf{br}(\mathcal{R})$ minus 1, i.e. $\mathsf{h}(\mathcal{R}) = |\mathsf{br}(\mathcal{R})| - 1$. It corresponds to the length of the path linking $\mathcal{R}$ to the root node. The height of the root node is set by convention to 0.*
- *The **subtree** rooted at $\mathcal{R}$, $\mathcal{T}(\mathcal{R})$, corresponds to all the elements of $\mathcal{T}$ that are included in $\mathcal{R}$. In other words, it contains all the elements of $\mathcal{T}$ for which $\mathcal{R}$ is an ancestor.*

1.3.3.2 Examples of tree-based image representation

Classical tree-based representation include the min-tree and max-tree, also known as *component trees*. Initially proposed in [173], these tree-based representations are based on threshold decomposition of a gray-scale image: they encodes the inclusion relationship between the connected components of the upper and lower level sets of the image. More specifically, let $\mathcal{I} : E \to V \subseteq \mathbb{R}$ be a gray-scale image. Its upper and lower level sets, for a threshold value

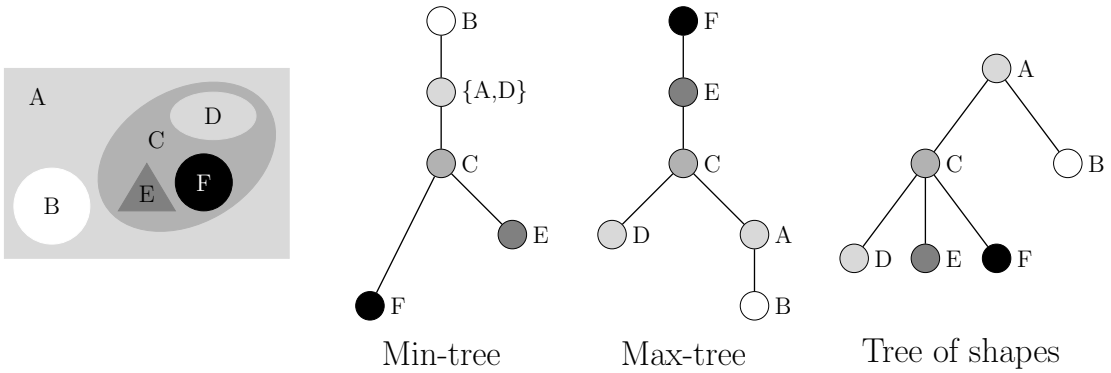


Figure 1.16: Examples of tree-based threshold decompositions of the gray-scale image on the left: min-tree, max-tree and tree of shapes.

h, are defined by:

$$\mathcal{I}^h = \{x \in E | \mathcal{I}(x) \geq h\} \tag{1.7}$$

$$\mathcal{I}_h = \{x \in E | \mathcal{I}(x) \leq h\} \tag{1.8}$$

$\mathcal{I}^h$ and $\mathcal{I}_h$ are binary images, composed of connected components, where each connected component C^h (respectively, C_h) corresponds to a set of connected pixels whose value is above (respectively, below) the threshold h. By varying the threshold h, one then obtain a hierarchical decomposition of the image into a set a connected components. The min-tree represents this hierarchical decomposition by encoding the inclusion relationship between the connected components of the lower level sets of the image. The leaves of the min-tree are the regional minima of the image. Conversely, the max-tree encodes the inclusion between the connected components of the upper level set decomposition, and has the local maxima as leaves. Examples of min-tree and max-tree are displayed by figure 1.16. Component trees have found numerous applications derived from mathematical morphology, such as filtering [97, 174] or segmentation [96] and several efficient implementations have been proposed (see [42] for a comparative review).

Despite their usefulness, components trees have several drawbacks. As a matter of fact, they handle bright and dark components separately. This can be an issue for instance when some object of interest appears brighter than the background in some parts of the image, and darker in some other parts. Moreover, real objects of interest may even not correspond to extrema of the image. Finally, the structure of components trees is bound to the pixel values since they must be comparable. While this works well for gray-scale (hence, scalar) images, it does not straightforwardly extend to multi-valued images where no natural order exists between vectors.

To handle bright and dark components in a self-dual way, several authors have introduced the notion of *shapes* which have led to the definition of the tree of shapes (ToS) (also called topographic map [45], or *inclusion tree* [140]). Instead of considering the connected components of the upper and lower level set decompositions, the ToS encodes the inclusion relationship

between the level lines (*i.e.*, the topological boundaries of the connected components). More particularly, a shape is defined as a connected component with holes filled, and the ToS of an image can be viewed as a merging between the min-tree and max-tree of this image. All leaves of the ToS correspond to some regional minima and maxima of the image, as shown by figure 1.16. In the same way as component trees, inclusion trees find applications in image filtering [229], segmentation [41], simplification [14] and object recognition [154]. However, similarly to component trees, an ordering on the pixel values is also required to built the ToS (since it is based on the notion of shape, itself deriving from the notion of connected component), hence making the extension from gray-scale to multi-valued image challenging. Several extensions have been proposed (see for instance [43, 44]), mainly based on the computation of marginal ToS (*i.e.*, one ToS per image channel) that are further merged.

1.3.3.3 Hierarchies of partitions

Component and inclusion trees are extrema-oriented image representations. They describe an image as a set of disjoint or nested connected components. However, they rely on the absolute pixel scalar values of the image and on the presence of a total ordering holding on this set of scalar values. Moreover, there is no guarantee that objects of interest can be appropriately described only by their own pixel values. As a matter of fact, an object seems to be of interest if it is sufficiently different from its surrounding. This leads to work on *dissimilarities* between pixels (or regions) rather than on their *absolute values*, in particular through the introduction of a dissimilarity function, which is notably the purpose of hierarchies of partitions.

As previously said, hierarchies of partitions are a special case of tree-based image representation. As a matter of fact, the definition 1.7 can be complemented to define a hierarchy of partitions, hereafter denoted H:

Definition 1.9 (Hierarchy of partitions - region-wise definition)
A hierarchy of partitions H of E is a collection of regions $H = \{\mathcal{R} \subseteq E\}$ such that:

$$\begin{aligned} &- \emptyset \neq H \\ &- E \in H \\ &- \forall \mathcal{R}_i, \mathcal{R}_j \in H, \mathcal{R}_i \cap \mathcal{R}_j \in \{\emptyset, \mathcal{R}_i, \mathcal{R}_j\} \\ &- \forall \mathcal{R} \in H \backslash \mathsf{leaves}(H), \mathcal{R} = \bigcup_{\mathcal{R}_c \in \mathsf{C}(\mathcal{R})} \mathcal{R}_c \end{aligned} \tag{1.9}$$

In addition to being composed of regions that are pairwise disjoint or nested, the additional requirement for a tree-based representation to be a hierarchy of partitions is that each non-leaf node in the hierarchy can be exactly recomposed from its children. In particular, it means that the whole space E can be retrieved by taking the union of all leaves of the hierarchy, which was clearly not the case for the component and inclusion trees (see figure 1.16). These leaf regions form a partition of E, denoted π_0 and that will be called the *leaf partition* of H.

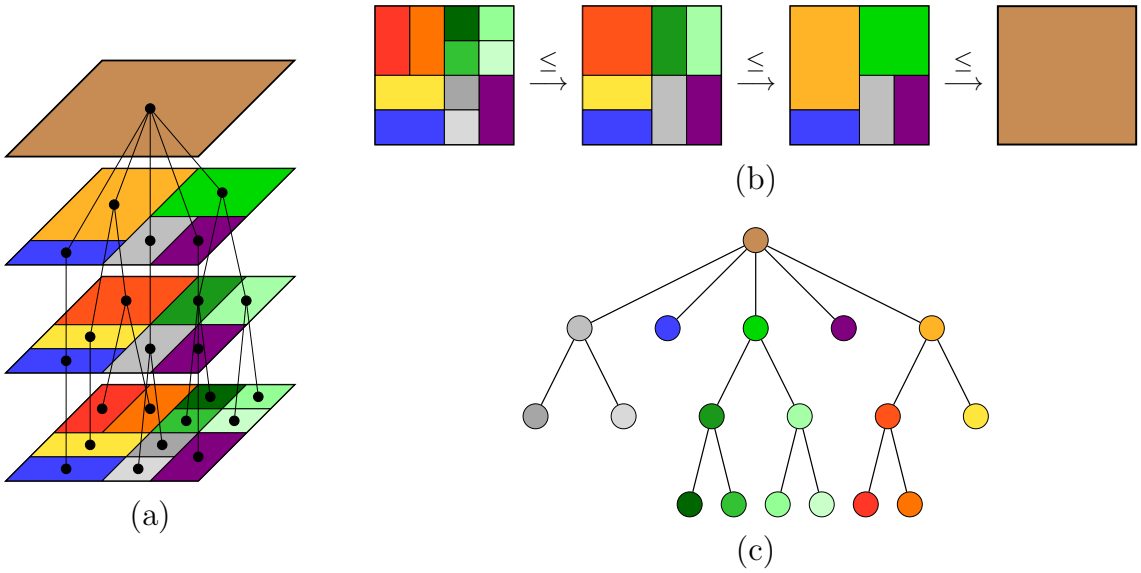


Figure 1.17: Example of hierarchy of partitions, viewed as (a) a stack, and (b) a sequence of partitions ordered by refinement, along with (c) the corresponding tree graph.

Alternatively, but equivalently to the definition 1.9, a hierarchy of partitions can be defined from a partitioning point of view:

Definition 1.10 (Hierarchy of partitions - partition-wise definition)
A hierarchy of partitions H of E is a finite sequence of partitions $\pi_i \in \Pi_E$ ordered by refinement:

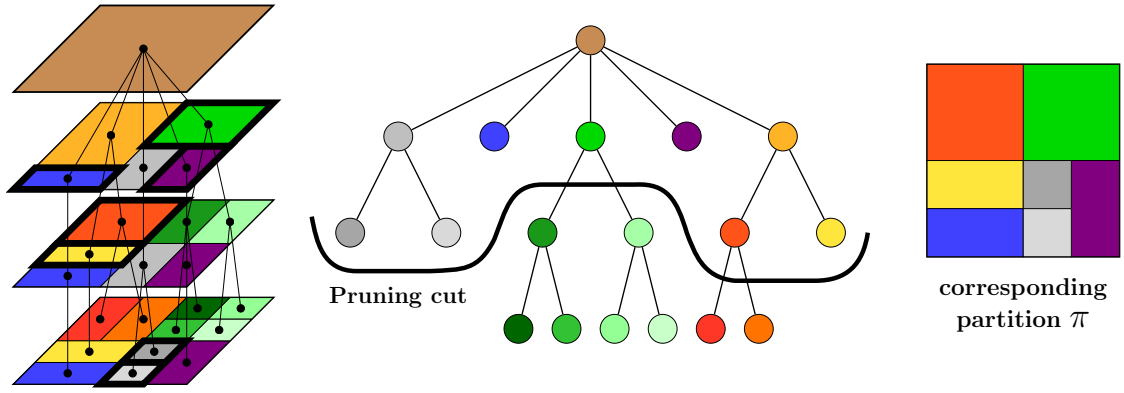
$$H = \{\pi_i\}_{i=0}^{n} \text{ such that } i \leq j \Rightarrow \pi_i \leq \pi_j. \tag{1.10}$$

The partitions are ranging from the leaf partition π_0 to the root of the hierarchy $\pi_n = \{E\}$. In definition 1.10, the word hierarchy takes its meaning since the partitions of the sequence are ordered from fine to coarse. An example of hierarchy of partitions and its associated tree graph is displayed by figure 1.17.

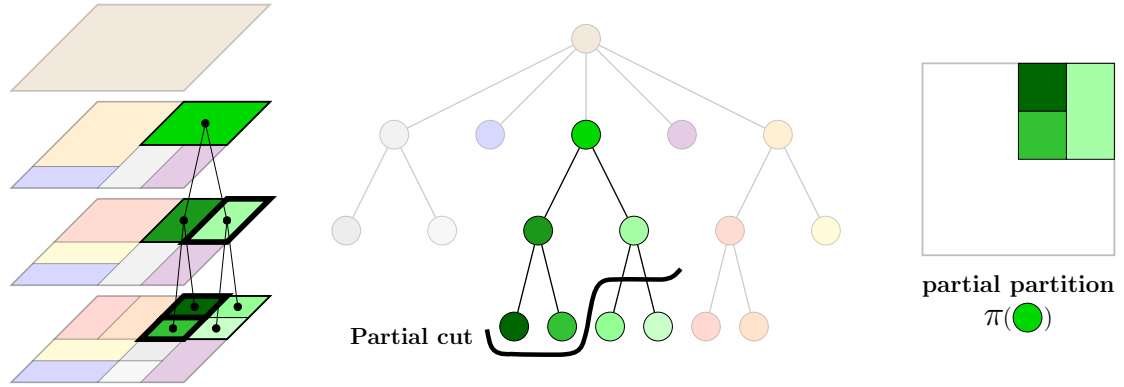
Thanks to these two equivalent definitions, it is possible to obtain a hierarchy either by working on the regions (for instance, using some region merging or splitting techniques) or on the partitions directly. Of course, the whole terminology 1.8 defined for tree-based representations remains valid for hierarchies of partitions.

Processing that are commonly applied to hierarchies of partitions can be categorized in two classes:

– Region-based processings. They aim at exploring the regions of the hierarchy in order to identify the regions of interest that fulfill some predefined criteria (for instance a given shape, homogeneity or distance with respect to the neighbors). These strategies are particularly useful to perform object detection and recognition, as it will be further investigated in chapter 3.



(a)



(b)

Figure 1.18: Example of (a) a cut of a hierarchy and its associated partition, and (b) a partial partition.

– Partition-based processings. Their goal is to extract from the hierarchy some specific partitions that conform a given application. One particular way to proceed is through a *pruning* operation, namely cutting of some branches of the hierarchy such that the new leaves of the pruned tree achieve the desired partition. Some pruning strategies will be investigated in chapter 2 and chapter 4.

Following is the related terminology:

Definition 1.11 (Cuts of a hierarchy)
A cut of a hierarchy H is a partition π of E whose regions belong to H. The set of all cuts of a hierarchy H of E is denoted $\Pi_E(H)$, and it is a sub-lattice of Π_E for the refinement ordering.

Definition 1.12 (Partial partition)
A partial partition $\pi(\mathcal{R})$ of $\mathcal{R} \in H$ is a cut of the sub-hierarchy $H(\mathcal{R})$. The support of this partition is only partial with respect to E, hence the name. As for the cuts, the set of all partial partitions of $\mathcal{R} \in H$ is denoted $\Pi_E(H(\mathcal{R}))$.

Graphically, a cut can be seen as a path that intersects each branch of the hierarchy at most once, as displayed by figure 1.18a. The regions constituting the corresponding partition are those whose associated node in the tree graph is located directly above the cut. Notice that, for some authors, the partition is made of the regions located below the cut. $\Pi_E(H)$ being a sub-lattice of Π_E means in particular that the supremum and infimum of two cuts of a hierarchy are also cuts of this hierarchy.

1.3.3.4 Examples of hierarchies of partitions

As for tree-based representations, hierarchies of partitions have been widely studied in the literature. A well-known hierarchy of partitions is the *quad-tree*, proposed by [75]. Starting from the whole image (*i.e.*, the root of the hierarchy), the quad-tree is created by successive region splitting. More particularly, each region, also called quadrant, can be either divided into four sub-quadrant or left as it is, each quadrant being either square or rectangular. The decision of splitting a region into four sub-quadrant is often based on some homogeneity considerations: if the region is not homogeneous enough, it is split until it fulfills the desired criterion. Quad-trees have found applications in image segmentation [184] and compression [182], notably. However, as each region is rectangular, quad-tree cannot account for irregular contours and therefore objects of interest are often split into several nodes.

Another type of hierarchy is the so-called α-*tree* [148], also known as the hierarchy of quasi flat zones [139], based on the notion of constrained connectivity [183]. More specifically, let p and q be two neighboring pixels, and $d(\mathcal{I}(x), \mathcal{I}(y))$ be the dissimilarity between their respective values for the image $\mathcal{I}$. Two pixels p and q are said to be α-connected if there is a path from p to q, namely a sequence of $(p = x_1, \ldots, x_n = q)$ such that x_i and x_{i+1} are adjacent (in the sense of the 4-neighbors or 8-neighbors systems) and $d(\mathcal{I}(x_i), \mathcal{I}(x_{i+1})) \leq \alpha$. Following, one can define the α-connected component of a pixel p (abbreviated $\alpha-\mathrm{CC}(p)$) as

$$\alpha - \mathrm{CC}(p) = \{p\} \cup \{q \text{ s.t } p \text{ and } q \text{ are } \alpha - \text{connected}\} \tag{1.11}$$

It was shown in [183] that for a given α value, the set of $\alpha-$CC forms a partition π_α of E and that, for two values $\alpha_1 \leq \alpha_2$, $\pi_{\alpha_1} \leq \pi_{\alpha_2}$. Therefore, by using several values $\alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_n$, one induces several partitions ordered by refinement that creates the $\alpha-$ tree hierarchy $H_\alpha = \{\pi_{\alpha_0} \leq \cdots \leq \pi_{\alpha_n}\}$. An example of such hierarchy is displayed by figure 1.19.

It is known that such defined $\alpha-$trees may suffer from the so-called *chaining effect*. For instance, a ramp image where all pixels of a given column have the same value, and such that the dissimilarity between two pixels of adjacent columns is equal to 1 would have a $1-$CC being the whole image. To tackle this issue, the $(\alpha, \omega)-$connected component of a pixel p was introduced [183] as the largest $\alpha-$CC containing p such that the maximum dissimilarity between two pixels of the connected component is less that ω:

$$(\alpha, \omega) - \mathrm{CC}(p) = \bigvee \{\alpha_i - \mathrm{CC}(p) \text{ s.t } \alpha_i \leq \alpha \text{ and } \max_{x,y \in \alpha_i - CC(p)} d(\mathcal{I}(x), \mathcal{I}(y)) \leq \omega\} \tag{1.12}$$

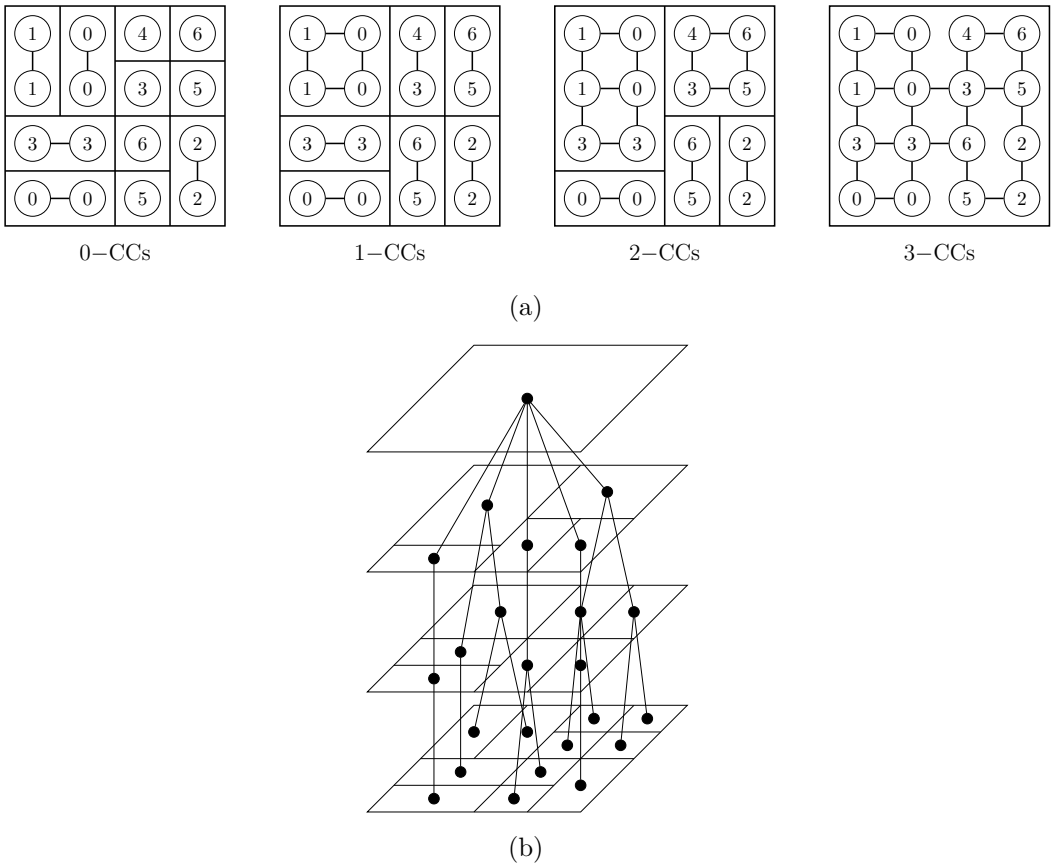


Figure 1.19: Example $\alpha-$tree hierarchy: (a) $0-, 1-, 2-$ and $3-$connected components of a toy image (with adjacency defined by 4-connectivity) and (b) the corresponding $\alpha-$tree. In that example, the dissimilarity measure between pixels p and q is $d(p,q) = |\mathcal{I}(p) - \mathcal{I}(q)|$.

Following this definition, $(\alpha_1, \omega_1)-\text{CC(p)} \subseteq (\alpha_2, \omega_2)-\text{CC(p)}$ for $\alpha_1 \leq \alpha_2$ and $\omega_1 \leq \omega_2$, and this also allows for the generation of fine to coarse partitions of E (and thus of a hierarchy) by progressively increasing the values of the range parameters α and ω. Some efficient algorithm to compute such hierarchies can be found in [143, 148].

Last but not least, a popular hierarchical representation is the *binary partition tree* (BPT), as proposed by [172]. Starting from an initial partition π_0 that defines the leaves of the hierarchy, the BPT is obtained by a bottom-up region merging procedure: pairs of neighboring regions are merged based on their similarity until there is only one region remaining, which is the whole space E. The creation of a BPT is bound to the definition of the initial partition as well as the similarity function to assess how close are two neighboring regions. In the last decade, BPTs have proved to be a valuable tool for hierarchical image representation thanks to the great flexibility of their construction and analysis processes. Consequently, they have found numerous applications in image and video processing such as image segmentation [172, 207], filtering [5], compression [172] as well as object detection [120, 218] and object tracking [153, 196]. The next section is devoted to a more detailed insight of BPTs, as they are going to play a

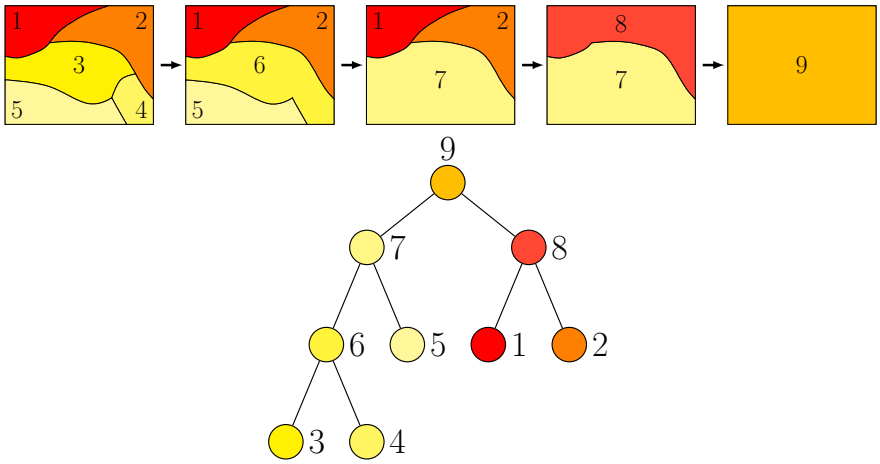


Figure 1.20: Example of region merging sequence along with its corresponding binary partition tree.

key role in the following chapters.

1.3.4 A focus on the binary partition tree

The binary partition tree encodes the hierarchical decomposition of an image in a tree structure. As said in the previous section 1.3.3.4, the BPT representation relies on the iterative merging procedure of a set of initial regions, which are the leaves of the BPT. The tree is built by keeping track of the merging orders. Each region can be merged with only one of its neighbors, resulting in a hierarchy where each region has either two children, or none (in the case of a leaf node). An example of region merging sequence and its corresponding BPT is displayed in figure 1.20. BPT representations enjoy several desirable properties:

- They allow to decompose an image into a set of regions that are hierarchically organized. This decomposition provides a description of the image at different scales ranging from fine to coarse. This is particularly valuable since the analysis of an image can be performed at different levels of details, according to the desired objectives. Therefore, the hierarchical decomposition can serve as an initial support, computed regardless of the application, and its analysis can be tuned afterward to meet the intended goal.
- The construction of the BPT is based on the merging of similar neighboring regions, and is therefore bound to the definition of this similarity measure. While this setting is left to the user and may appear at a first sight as a disadvantage with respect to strategies exploiting the absolute pixel values (such as the component and inclusion trees for example, which totally rely on the notion of regional extrema), it can actually be seen as a strength as it introduces some flexibility in the construction of the BPT.
- Even though their construction is rendered flexible by the various possible settings to parametrize the merging procedure, BPTs were intended to be built independently of the underlying application, as a common support basis for all subsequent processing [172].

The analysis of the BPT, which is driven by the application, can adapt well to a broad range of processing. For this reason, BPTs have been used in an extensive variety of applications in the image and video processing fields [6, 153, 207, 218].

1.3.4.1 Construction of the BPT

There are two parameters that are of prime importance when building a binary partition tree, namely the definition of the merging procedure, and the initial partition on which this procedure is applied. While there are various options available for those two parameters, some of them have proved to perform consistently well in the literature.

The merging procedure The merging procedure determines in which order the regions should be merged. The BPT is then built following a bottom-up procedure (*i.e.*, starting from the smallest regions) by keeping track of this order. The specification of a merging procedure itself relies on the definition of two inner parameters:

- The *region model* $\mathcal{M}_\mathcal{R}$, which specifies how to mathematically model the regions and their union.
- The *merging criterion* $\mathcal{O}(\mathcal{R}_i, \mathcal{R}_j)$, which assesses the similarity between neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ by measuring the distance between their region models $d(\mathcal{M}_{\mathcal{R}_i}, \mathcal{M}_{\mathcal{R}_j})$.

Relevant definitions for the region model and its associated merging criterion with respect to the processed image should guarantee the consistency of its BPT representation.

BPT were initially developed in the scope of gray-scale and color image processing [172], that is, for images whose space of pixel values V is either $\mathbb{R}$ or $\mathbb{R}^3$. In that case, the first proposed region model was the mean color within the each region:

$$\mathcal{M}_\mathcal{R} = \boldsymbol{\mu}_\mathcal{R} = \frac{1}{|\mathcal{R}|} \sum_{x \in \mathcal{R}} \mathcal{I}(x) \tag{1.13}$$

with $\mathcal{I}(x)$ being a scalar (respectively, a triplet) in the case of gray-scale (respectively, color) images. This model, assuming color homogeneity within the region, allows the use of simple merging criteria and can be easily computed for a node given the regions models of its children, thus leading to fast and efficient implementations. While some authors use it directly on the common RGB color space [20], it is more often applied on other spaces, such as the LUV [172] or the CIE $L^*a^*b^*$ [120, 127] color spaces, which are known to better match the human visual perception in terms of distance between colors. All the previously cited works used the following merging criterion, introduced in [78], to measure the similarity between neighboring regions $\mathcal{R}_1$ and $\mathcal{R}_2$

$$\mathcal{O}(\mathcal{R}_1, \mathcal{R}_2) = |\mathcal{R}_1| \times \|\mathcal{M}_{\mathcal{R}_1} - \mathcal{M}_{\mathcal{R}_1 \cup \mathcal{R}_2}\|_2 + |\mathcal{R}_2| \times \|\mathcal{M}_{\mathcal{R}_2} - \mathcal{M}_{\mathcal{R}_1 \cup \mathcal{R}_2}\|_2 \tag{1.14}$$

with $\|\cdot\|_2$ being the Eucliden L_2 norm. Other norms, such as the L_1 and the L_∞ norms can also be used (recall that the L_p norm of a vector $\mathbf{x}$ for $p \geq 1$ is defined as $\|\mathbf{x}\|_p = (|x(1)|^p + \cdots + |x(n)|^p)^{\frac{1}{p}}$).

BPT representations were extended to hyperspectral images in the work of Valero [204,205, 207,210]. In such case, the space of pixel values V is $\mathbb{R}^N$, with N typically equal to several hundreds. The extension of the mean region model defined by equation (1.13) to a larger dimensionality N is straightforward, and is termed *mean spectrum* or *first order* region model, following [205]. However, the design of suitable merging criteria for this region model is needed, since it is known that L_p norms suffer from the curse of dimensionality. To alleviate this issue, Valero proposed two merging merging criteria, adapted to the inherent large dimensionality of hyperspectral images:

– The *spectral angle* (often abbreviated SAM) between two regions $\mathcal{R}_i$ and $\mathcal{R}_j$ is defined as the angle between their mean spectrum $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$:

$$\mathcal{O}_{SAM}(\mathcal{R}_i, \mathcal{R}_j) = \arccos\left(\frac{\boldsymbol{\mu}_i^T \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i\|_2 \|\boldsymbol{\mu}_j\|_2}\right) \tag{1.15}$$

This merging criterion is motivated in hyperspectral imagery by the fact that two spectrum describing the same material should have similar shapes, and thus a small angle between them in the feature space. Note also that the SAM is relatively insensitive to scaling effect since multiplying a vector by a constant only changes its magnitude, but not its angle.

– The *spectral information divergence* (SID) measure the distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ when interpreted as probability density functions. As a matter of fact, if a mean spectrum $\boldsymbol{\mu}$ is normalized to sum up to one ($\boldsymbol{\mu}^\star = \boldsymbol{\mu}/(\mathbb{1}^T\boldsymbol{\mu})$, with $\mathbb{1}$ being a column vector of ones), it can then be viewed as a probability density function. A common measure of similarity between such probability density functions is the so-called Kullback-Leibler divergence

$$d_{KL}(\boldsymbol{\mu}_i^\star, \boldsymbol{\mu}_j^\star) = \sum_{k=1}^{N} \mu_i^\star(k) \log\left(\frac{\mu_i^\star(k)}{\mu_j^\star(k)}\right) \ . \tag{1.16}$$

$d_{KL}(\boldsymbol{\mu}_i^\star, \boldsymbol{\mu}_j^\star) \geq 0$, and the equality is reached if and only if the two probability density functions coincide. However, as this measure is not symmetric (as it is a divergence and not a distance), the SID merging criterion is defined as the symmetric Kullback-Leibler divergence between $\boldsymbol{\mu}_i^\star$ and $\boldsymbol{\mu}_j^\star$:

$$\mathcal{O}_{SID}(\mathcal{R}_i, \mathcal{R}_j) = d_{KL}(\boldsymbol{\mu}_i^\star, \boldsymbol{\mu}_j^\star) + d_{KL}(\boldsymbol{\mu}_j^\star, \boldsymbol{\mu}_i^\star) \ . \tag{1.17}$$

As for traditional images, the first order region model is simple and assumes spectral homogeneity within the region. However, this may become a limitation for some applications where the spectral variability has to be taken into account. To that purpose, the *non-parametric statistical* region model, also called *histogram-based* region model, was introduced [205]. This model is defined as a set of N histograms:

$$\mathcal{M}_\mathcal{R} = \{\mathcal{H}_\mathcal{R}^1, \ldots, \mathcal{H}_\mathcal{R}^N\} \tag{1.18}$$

where each $\mathcal{H}_\mathcal{R}^i$ is the empirical spatial distribution of the pixel values within region $\mathcal{R}$ for the i^{th} band. More particularly, each histogram $\mathcal{H}_\mathcal{R}^i$ is composed of N_{bins} bins $a_p, p = 1, \ldots, N_{bins}$:

$$\mathcal{H}_\mathcal{R}^i = \{\mathcal{H}_\mathcal{R}^i(a_1), \ldots, \mathcal{H}_\mathcal{R}^i(a_{N_{bins}})\} \tag{1.19}$$

and each histogram, if normalized to sum to one, can also be viewed as an approximation of the probability density function. Note that this region model can also be recursively computed for each region $\mathcal{R}$ as the weighted sum of the regions models of its children. This histogram-based region model allows to define for merging criteria some metrics that measure the similarity between histograms. In particular, Valero [205,207] introduced three histogram-based merging criteria:

– The *Battacharyya distance*, which is based on the Battacharyya coefficient (BC) between two normalized histograms $\mathcal{H}^i_{\mathcal{R}_1}$ and $\mathcal{H}^i_{\mathcal{R}_2}$ of two adjacent regions $\mathcal{R}_1$ and $\mathcal{R}_2$, for the same band i:

$$BC(\mathcal{H}^i_{\mathcal{R}_1}, \mathcal{H}^i_{\mathcal{R}_2}) = -\log\left(\sum_{p=1}^{N_{bins}} \sqrt{\mathcal{H}^i_{\mathcal{R}_1}(a_p)}\sqrt{\mathcal{H}^i_{\mathcal{R}_2}(a_p)}\right) \tag{1.20}$$

If the two histograms perfectly overlap, the argument within the logarithm sums to one, hence a Battacharyya coefficient being 0. Consequently, the merging criterion based on the Battacharyya distance can be obtained by summing the Battacharyya coefficients for all the N bands of the image:

$$\mathcal{O}_{BC}(\mathcal{R}_i, \mathcal{R}_j) = \sum_{i=1}^{N} BC(\mathcal{H}^i_{\mathcal{R}_1}, \mathcal{H}^i_{\mathcal{R}_2}) \tag{1.21}$$

The main limitation of this merging criterion is its assumption that the histograms are perfectly aligned, hence its name of *bin-to-bin* distance. This can be a disadvantage in a situation where two histograms have a similar profile but are not aligned, and one may want to consider those two histograms as close to each other.

– The *diffusion distance*, proposed in [119] and which solves the previously raised issue concerning two histograms that do not overlap. For that reason, the diffusion distance is called a *cross-bin* distance. It is based on the idea that the difference between two histograms

$$d^i_0(a_p) = \mathcal{H}^i_{\mathcal{R}_1}(a_p) - \mathcal{H}^i_{\mathcal{R}_2}(a_p), \quad p = 1, \ldots, N_{bin} \tag{1.22}$$

can be viewed as a temperature field, and the corresponding distance between those two histograms is the time needed by this field to reach stability via a heat diffusion process, or equivalently, on the state of the temperature field after a given time. More precisely, starting from d_0, the diffusion process is simulated by convolving the current temperature distribution with a Gaussian kernel

$$d^i_m(a_p) = [d_{m-1}(a_p) * g_\sigma(a_p)] \downarrow_2, m = 1, \ldots, M \tag{1.23}$$

with g_σ standing for the Gaussian kernel with variance σ, $\downarrow_2$ denotes a downsampling by a factor of 2, and M is the number of convolution layers. The final merging criterion between $\mathcal{R}_1$ and $\mathcal{R}_2$ follows by summing over all N bands the L_1 norm of the $M+1$ layers of temperature fields

$$\mathcal{O}_{DIF}(\mathcal{R}_1, \mathcal{R}_2) = \sum_{i=1}^{N}\sum_{m=0}^{M} \|d^i_m\|_1 \tag{1.24}$$

Similarly to the Battacharyya distance, the diffusion distance processes all bands of the image in the same way. On the other hand, hyperspectral data feature strong correlations between bands, and this correlation could be used to remove the redundant information contained in each region model.

– The *similarity via multidimensional scaling* aims exactly at exploiting the redundancy between all bands of the hyperspectral image. First, a multidimensional scaling [55] is performed on the N histograms of $\mathcal{M}_{\mathcal{R}}$ in order to reduce the dimensionality and extract only the principal components of the region containing the most relevant information. Then, for any two neighboring regions $\mathcal{R}_1$ and $\mathcal{R}_2$, the similarity measure between those two regions is obtained by analyzing the joint correlation between the principal components of each region. More specifically, a statistical test, based on a multivariate analysis of variance (MANOVA) [9] is performed in order to determine whether the principal axes are correlated or not. In the first case, a dependency is claimed between regions $\mathcal{R}_1$ and $\mathcal{R}_2$, which are thus given a low distance. Details about the implementation of this merging criterion can be found in [204, 207].

While it is appealing to be able to define several region models and their non-exhaustive list of merging criteria, it also raises the question on which couple $(\mathcal{M}_{\mathcal{R}}, \mathcal{O}(\mathcal{R}_i, \mathcal{R}_j))$ would lead to the most consistent hierarchical representation given an image. Although there is no clear answer to this question, we can formulate this heuristic rule, supported by the similar conclusions drawn in the PhD work of Valero [204]:

– If one is interested by relatively simple and spectrally homogeneous regions, then the mean spectrum region model is a good candidate. Provided this region model, the SAM and SID merging criteria perform equally well.

– Alternatively, if one is giving importance to the intra-region spectral variability, then one should choose the histogram-based region model, which is however computationally heavier than its mean spectrum counterpart. Related to the merging criteria, the diffusion distance performs better than the Battacharyya since it is a cross-bin distance. The similarity via multidimensional scaling in turn gives more consistent results than the diffusion distance since it takes into account the correlation between bands of the hyperspectral image, but at the cost of a higher computational burden.

In both cases, the region size does not intervene in the previously defined merging criteria, and this could lead to small and insignificant regions remaining in the last merging iterations of the construction. To overcome this issue, it was proposed in [36] to use a priority rule: all regions whose size is less than a given threshold (typically set to 15%) of the mean size of the regions standing in the current merging iteration are given the merging priority, regardless of their distance with respect to their neighbors.

The initial partition The second parameter needed for the BPT construction is the initial partition π_0, on which is initialized the region merging procedure. If a pertinent initial partition does not guarantee a pertinent BPT representation (since it also depends on the definition of the region model and merging criterion), a poor initial partition does lead to a poor hierarchical decomposition, as all the regions subsequently obtained follow from the

(a) (b) (c)

Figure 1.21: Examples of initial partitions with (a) the multidimensional watershed, (b) the mean shift clustering and (c) SLIC superpixels.

initial ones. The safest option that could be considered is to initialize the BPT construction at the pixel level (*i.e.*, where the initial partition is composed of regions made of single pixels only). Since a BPT built on an initial partition made of $|\pi_0|$ leaves is composed of $2|\pi_0| - 1$ regions, the pixel level as an initial partition may lead to a huge BPT structure and this could be problematic from a computational point of view for very large images. Moreover, such BPT would be composed of a lot of small and meaningless regions and this could also slow down the analysis processes further applied on it.

Then, an appropriate initial partition should enjoy the following two properties:
- Its regions should be fine enough (in other words, the image should be enough over-segmented) to ensure that the smallest regions of interest within the image are not already merged together in some initial regions. Otherwise, those regions of interest would be irremediably lost.
- The boundaries of the initial regions should well adhere to the real boundaries of objects of interest, in order to be able to reconstruct (up to a correct definition of a region model and merging criterion) these objects of interest accurately.

If those two conditions are fulfilled, it was shown (in a context of image segmentation) that starting from an initial partition does not worsen the segmentation results [199].

Among efficient segmentation algorithms to design the initial partition, one can cite the watershed algorithm [220] (or the multidimensional watershed for multi-valued images, see [188]), the mean shift clustering [52] or the SLIC superpixels [1]. All those fulfill the two conditions of over-segmentation and boundary adherence, as it can be seen in figure 1.21.

1.3.4.2 Processing of the BPT

Once its construction is completed, the BPT encodes in its structure a decomposition of the image in regions at various scales, from the finest ones (*i.e.*, the leaves) to the coarsest

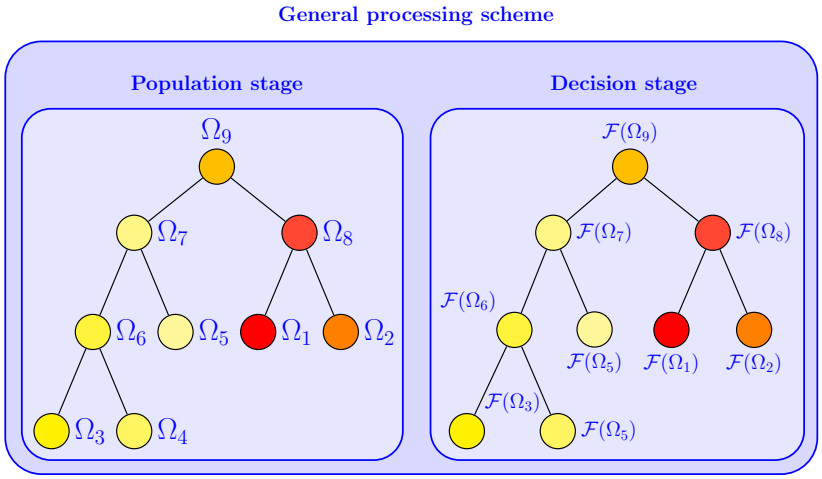


Figure 1.22: General scheme of a the BPT processing step

(the root of the hierarchy being the whole image). Contrarily to the construction, which can be done regardless of this application, the strategy to further process this set of hierarchically organized regions is strongly driven by the underlying application. As a matter of fact, one is not going to operate the same way whether one wants to extract a particular cut from the hierarchy (for segmentation purpose for instance) or one seeks a particular object of interest in the image (for example in a context of object detection).

Nevertheless, a typical BPT processing can be decomposed in two steps: the population of the tree in a first stage and a following decision stage, which are both defined to achieve the intended goal. During the former, some features or attributes are evaluated for each region $\mathcal{R}$ and stored in a set $\Omega_{\mathcal{R}}$: the tree is "populated". Then, the decision step evaluates, given a decision rule, if each node should be retained or discarded according to its previously computed set of features. This decision step involves a decision function $\mathcal{F}$ that is applied on each region to take the decision whether to keep this node or not. This whole scheme is illustrated in figure 1.22.

As a simple example to illustrate this processing chain, consider an application where one wants to smooth an image by removing small and inhomogeneous regions. A possible strategy to achieve such goal would be to filter out from the BPT representation all regions whose size $|\mathcal{R}|$ is below a predefined threshold δ. As a consequence, the attribute that should be retrieved for each region is its area, defining a feature set $\Omega_{\mathcal{R}} = \{|\mathcal{R}|\}$. The decision rule being *remove a node if its size is below the threshold*, the decision function then becomes $\mathcal{F}(\Omega_{\mathcal{R}}) \overset{?}{\geq} \delta$, and all nodes which do not satisfy this decision are removed from the BPT, producing a pruned tree where all new leaves have a size greater than or equal to the size threshold δ.

This decision function, applied on the region area, is a particular case of increasing decision: a decision is said to be increasing if $\mathcal{R}_1 \subseteq \mathcal{R}_2 \Rightarrow \mathcal{F}(\Omega_{\mathcal{R}_1}) \leq \mathcal{F}(\Omega_{\mathcal{R}_2})$. In such case, if a node has to be retained, then so have to be all its ancestors. Conversely, if it is decided that a

node should be discarded, it is also the case for all its descendants. When the decision is not increasing, then some more sophisticated strategies have to be used, such as the *minimum*, *maximum*, or *Viterbi* decision rules. The *minimum* decision rule states that a region is preserved if and only if all its ancestors also have to be preserved. The *maximum* decision is the opposite, namely a node is removed if and only if all its descendant also have to be removed. The *Viterbi* decision strategy, on the other hand, associates to each node a cost reflecting how much impact it would have to change the decision for this node (for instance, how much would it cost to remove a node that was decided to be retained). It then tries to minimize this cost function in order to make the decision function increasing. As an example, if it has been decided that all nodes in a whole branch should be retained except for one, it is less costly to take the decision to retain all nodes of the branch (so inverting the decision for only one node) rather than removing all nodes (see [172, 204] for more details).

1.3.5 Conclusion on hierarchical representations

In this section we have presented the concept of hierarchical representations of images, which are a particular case of tree-based image decompositions. Tree-based image decompositions naturally arise in image processing because natural images can often be decomposed in a set of regions of interest (which our brain can interpret with a semantic meaning) which are organized in a hierarchical manner, from fine to coarse. In addition, an image can be analyzed at various levels of details, which is driven by the application and the information one expects to extract from it. Tree-based image decompositions allow to compile in their tree structure all the potential scales of interest. The decomposition can then be computed once for an image, and its further analysis is tuned in accordance with the goal to achieve. Tree-based image representations find numerous applications in image processing, such as image segmentation, filtering or object detection.

Several tree-based image decompositions have been proposed in the literature, the most popular being the component and inclusion trees. However, their constructions rely on the ordering by inclusion of regional minima and maxima, which is rendered possible when handling gray-scale images because pixel values are scalar and can be easily compared. When dealing with multi-valued images, one achieves the comparison of pixel values by introducing some dissimilarity measure, which is at the core of the definition of a hierarchy of partition. Notable hierarchies include the quad-tree, the $\alpha-$tree and the binary partition tree.

In particular, we focused more in details on the binary partition tree. Given an initial partition of the image and a bottom-up region merging algorithm, one obtains a BPT representation by merging iteratively neighboring regions based on their similarity, until only one region remains (which is the whole image support). The description of a proper region merging algorithm requires the definition of a region model, *i.e.*, a mathematical formulation to model a region, and a merging criterion, which measures the similarity between two neighboring regions by computing the distance between their respective region models. BPTs have received much attention lately, by their capacity to handle images with very high dimensionality, such as hyperspectral images (which contain up to several hundreds of spectral channels), and they

are now considered as a standard image processing tool.

1.4 Example of a BPT-based application

We have developed so far the key aspects to properly operate the BPT representation in order to achieve a given goal, namely how to construct the BPT in order to obtain a consistent hierarchical decomposition of the image, and how to process it properly in order to achieve a given application.

The goal of this section is to demonstrate with a practical example all the questions that one has to answer to make the most of a BPT representation (and more generally, a hierarchical representation), namely how to properly design the construction and subsequent processing steps. BPT were initially proposed as a support basis for hierarchical analysis of images, and should therefore not be built to suit one particular application rather than another. Nevertheless, the definition of the initial partition, the region model and the associated merging criterion should be done in accordance with the specificities of the image as they directly impact the consistence of the hierarchical decomposition. The desired goal will then be taken into account when designing the analysis process of the resulting BPT.

More particularly, we focus on the segmentation of a tropical rain forest hyperspectral image. This application has been discussed in our previous work [199], from which we summarize the main points here.

1.4.1 The data set

The hyperspectral image analyzed here was captured over the Nanawale Forest Reserve, Hawaii (USA). The Nanawale forest is classified as lowland humid tropical forest, with an average elevation of 150 m above sea level. Mean annual precipitation and temperature are 3200 $mm.yr^{-1}$ and 23°C, respectively. The forest canopy is comprised of about 17 species, mostly invasive non-native trees, with a few native species remaining. The data were acquired with the Carnegie Airborne Observatory (CAO) Alpha sensor package in September 2007 [11]. The CAO-Alpha is equipped with a spectroscopic imager measuring up to 72 bands in the visible and near infrared domains. The collected hyperspectral image is composed of 1980×1420 pixels with 0.56 m ground sampling distance, covering an area of about 70 hectares on the ground. The spectral resolution used for this campaign resulted in the acquisition of 24 spectral bands of 28 nm in width and evenly spaced between 390 nm and 1044 nm. The whole 1980×1420 image contains several outlier pixels, as well as different flight lines. Therefore, for the purpose of this example only, we simply consider a 850×950 sub-image of the full data, displayed by figure 1.23.

Figure 1.23: Color composition of the Nanawale tropical rainforest hyperspectral image. Red, green and blue bands are centered on 646 nm, 561 nm and 447 nm respectively.

1.4.2 Construction of the BPT

In order to obtain the most consistent hierarchical representation, the various parameters such the the initial partition as well as the region model and merging criterion should be defined to make the most of the specificities of the image. In the present case, we aim at analyzing a hyperspectral image which was acquired over a forested cover. It means in particular that the overall spectral variability within the whole image is expected to be low, since all spectra depict the typical response of a tree. It is known that the global response of a tree to the incident light features a hump between 500 nm and 550 nm, whose height is due to the amount of chlorophyll contained in the leaves of the tree, a sharp rise at the edge of the near-infrared region (circa 700 nm) and then a drop around 1000 nm, due to the leaves water content. Even if two different tree species have their own particular signatures and proper features, their response should have the similar overall shape. This effect can be observed in figure 1.24. In addition, one can also expect some variability within the spectra corresponding to the same tree species, as this signal is also influenced by factors related to the foliage structure (such as the leaf angle distribution). One can see in particular in figure 1.24 the difference between two spectra belonging to the same species (see [74] for more details). In particular, this implies that the mean spectrum as region model for the construction of the BPT should be avoided, as it does not take into account the possible spectral variability within each region it models. It is then suggested *to use the histogram-based region model* instead.

Defining a proper region model is the first step toward a relevant BPT representation. The further point to analyze is the definition of an appropriate merging criterion. Considering that a tree crown can be partially shaded suggests that a cross-bin distance should be used

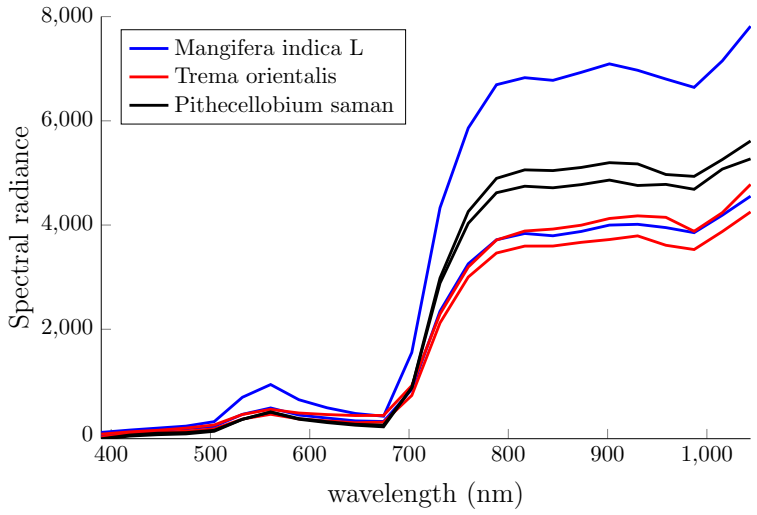


Figure 1.24: Example of tree spectra, where color corresponds to a particular tree species. The within-species variability can be appreciated, particularly for the bue spectra.

instead of a bin-to-bin one, thus excluding the Bhattacharyya distance. It is known that forested hyperspectral images feature strongly correlated bands, and this information should be also taken into consideration in order to choose the most adapted merging criterion. In particular, constructing the BPT over the principal components (PCs) of the image rather than over the raw hyperspectral bands can be considered as a potential solution. As a matter of fact, a principal component analysis (PCA) allows to uncorrelate the hyperspectral channels, such that each PC features the projection of the hyperspectral bands onto a particular factor influencing the spectra. For instance, the impact of brightness is particularly strong on radiometric signals measured from vegetation, and this influence is going to be expressed by the first PC (which resembles a gray-scale version of the hyperspectral image, see [199] for an example). The following PCs express features related to leaf chemistry (for instance, photosynthetic pigments or water content) and vegetation structure (foliage density), and those should help discriminating between various tree species.

As a matter of fact, if one can select the relevant PCs that contain discriminant information prior to the construction of the BPT, one should improve the ability of this BPT to differentiate trees belonging to different species. By conducting a visual analysis over the first few PCs, we came to the conclusion in [199] that only PC#2 to PC#8 contained some useful discriminant information and should be retained. Therefore, instead of selecting the similarity via multidimensional scaling merging criterion, which would have performed a similar analysis for each pair of regions during the construction of the BPT, the PCA transformation is performed once prior to the construction and *the diffusion distance is chosen*. It allows in addition to relieve the computation burden.

The final input parameter to define in order to built the BPT is the initial partition. In [199], we compared the multidimensional watershed and the mean shift clustering, and

showed that the latter was leading to better results. Therefore, we choose again in this example to derived the *initial partition with the mean shift clustering* algorithm [52]. Spectral and spatial bandwidths are both set to 5, producing an initial segmentation composed of 21 770 regions (to contrast with the potential $850 \times 950 = 807\ 500$ initial regions if the BPT was built over the pixel level).

1.4.3 Analysis of the BPT

The construction of the BPT has been defined to make the most of the image characteristics, in order to produce the most consistent hierarchical decomposition of the image. Its further analysis is however totally driven by the application. In our example, we aim at separating the various tree crowns in the image. Thus, we place ourselves in a segmentation context, which translates in terms of BPT processing as a pruning operation. We seek the best pruning cut π among all possible cuts $\Pi_E(H)$, with H being the BPT representation, and E being the image support.

Image segmentation is in itself an ill-posed problem, as a given image possesses as many acceptable segmentations as the number of possible applications for this image. In terms of BPT pruning strategy for the tree crown segmentation application, it suggests that a pruning strategy specifically dedicated to this goal would probably perform better than a more generic one. In [199], we proposed a pruning strategy based on the evolution of the region size along a branch. As remarked in [126], the evolution behavior of certain quantities along a full branch of the BPT provides some important information about the features contained in the image. In our case in particular, assuming that the initial partition is over-segmenting enough the image, it is possible to detect which regions in the BPT representation correspond to real tree crowns. In fact, each tree crown is over-segmented at first, and thus splits into several leaves. During the first iterations of the merging process, all leaves that are sufficiently close are going to be merged, and those leaves are assumed to belong to the same tree crown. At some point, all the leaves corresponding to a given tree crown will have merged into a bigger region, which will come to a steady state as it should lie farther apart from its neighbors. In the late steps of the merging process, this region will be forced to merge again, but its sibling at this time should also be a grown-up region. Therefore, when looking at the evolution of the region size along a branch, from a leaf to the root of the BPT, one should see a clear discontinuity at this stage where the region was forced to merge with a grown-up region in its surrounding. In [199], we remarked that the region prior to this discontinuity in the branch was the most likely to correspond to a tree crown.

Therefore, we designed a pruning strategy based on this observation, and following the same voting process scheme as presented in [206]. More particularly, each leaf of the BPT has its size evolution curve analyzed along its cooresponding branch. Given a size threshold δ, each leaf then votes for the region located prior to the first discontinuity in the branch, namely when the gap between the size of a node and its father along the branch exceeds δ.

(a) (b)

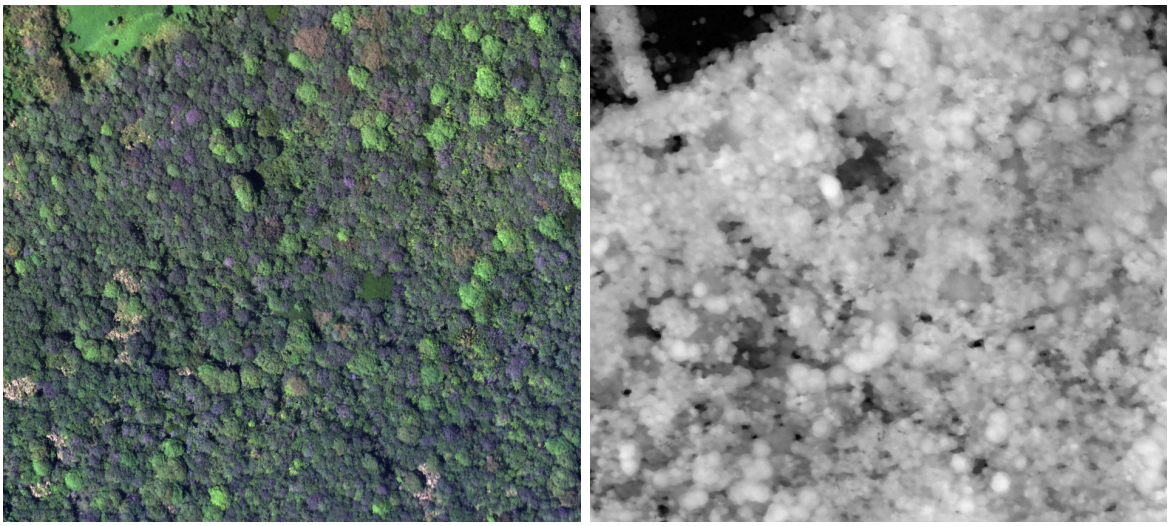
Figure 1.25: (a) Result of the proposed BPT-based segmentation, where each region is filled with the mean color of the original image (b).

Each node $\mathcal{R}$ has then its ratio

$$\frac{|\text{vote}(\mathcal{R})|}{|\mathsf{leaves}(\mathcal{R})|} \tag{1.25}$$

evaluated, where the numerator and the denominator are the received number of votes and number of leaves of $\mathcal{R}$, respectively. The decision rule which is then undertaken is to keep a node if it has a ratio vote/leaves of at least 1/2, namely if at least half of his leaves have decided to be represented by it. A maximum decision is finally conducted: a node is removed from the BPT if and only if all its descendant can be removed as well, leading to a pruned tree whose leaves define the desired segmentation. Figure 1.25 shows the result of this pruning strategy, applied to the tropical rain forest hyperspectral image, with a size threshold δ set to 2000. As can be seen, most of the tree crowns have been properly segmented.

In [199], we conducted a quantitative analysis based on some partial ground-truth data, where some reference tree crowns had been delineated by a trained operator. In particular, we compared the results of the proposed strategy against the results obtained by the pruning strategy proposed in [204, 206], which is based on a recursive spectral graph partitioning method and which can be considered as generic since it relies only on dissimilarities among nodes of the BPT and does not assume any particular knowledge about the currently processed image. We obtained up to 54.4% of properly segmented tree crowns for this data set with the proposed method, outperforming the recursive graph cut partitioning which correctly delineated 42.5% of the reference tree crowns. While those segmentation number may seem low, we recall that tropical rain forests are among the richest and most complex ecosystems in the world. Given the density of the canopy in terms of individuals and species, as well of the complexity of its structure, achieving a perfect delineation of each tree crown is highly unrealistic. However, even partial information allowing a better delimitation, identification and enumeration of certain species of interest (such that dominant, rare or invasive species that are key indicators for environmental processes) can help ecologists to better understand



(a) (b)

Figure 1.26: (a) Hyperspectral and (b) LiDAR modalities of the Nanawale tropical rain forest. The LiDAR image has the same 0.56 m ground sampling resolution as the hyperspectral image.

these complex ecosystems. Our proposed method is, to the best of our knowledge, the first reference study for the segmentation of tropical rain forest tree crowns. A segmentation method for hyperspectral images was developed in [35], and applied on Compact Airborne Spectrographic Imager (CASI) data acquired over mixed Australian forests. They reported over 70% of success for the segmentation of trees or clusters of trees belonging to the same species, for relatively sparse vegetation covers. However, they noted a significant drop in this segmentation accuracy for dense and complex canopies, which is consistent with our reported results.

1.4.4 The benefits and challenges of multimodality

The tree crown segmentation is also a way to illustrate how the use of multimodal information could be beneficial in a concrete. As a matter of fact, one can see when analyzing figure 1.25 that neighboring trees are often aggregated together in the final segmentation map when they belong to the same species. This pattern is due to the fact that the BPT representation of the image is solely relying on the spectral characteristics of the scene. Therefore, adjacent trees of the same species are likely to be represented by a single region in the final segmentation map. A possible solution to overcome this issue would be the joint use of hyperspectral and LiDAR data. Indeed, the integration of the height and the physical shape of the tree crowns, carried by the LiDAR modality, could help discriminating the case where several trees sharing the same spectral properties stand next to each other.

The use of LiDAR data for the tree crown delineation has already been thoroughly investigated in the literature, and several techniques were developed to make the most of its

specificities when acquired over forested covers. In particular, methods such as region growing, valley following, template matching or stochastic point processes have been proposed (see [199] and references therein). They proved to perform well on images of temperate forested areas (such as coniferous or deciduous forests) thanks to the regular and elongated shape of the tree stands and the rather sparse canopy limiting the overlap between neighboring individuals. Their performances significantly drop however when applied to tropical forest ecosystems where tree size and shapes are highly variable and trees usually overlap due to the dense canopy structure. Looking at figure 1.26, one can indeed see that the height and shape features are less discriminative, as it seems more challenging to accurately recognize tree crowns in the LiDAR modality (figure 1.26b) with respect to the hyperspectral modality (figure 1.26a).

Fusing somehow the information extracted from both the LiDAR and hyperspectral modalities is a promising direction of research, but it also raises several questions on the way this integration should be done. For instance, at which stage of the hierarchical analysis should this integration be done? Which relative weight should be given to the information provided by the LiDAR when it conflicts the hyperspectral modality (in the case of two neighboring trees of different species but of same height for instance)? This illustrates, if necessary, the kind of issues that have to be taken up to integrate and make the most of multimodal information in order to boost the performances of classical image processing and analysis tools.

1.5 Conclusion

In this first chapter, we have presented the two notions that are the cornerstones of this manuscript, namely multimodality and hierarchical representations. The concept of multimodal data reveals a huge potential when it comes to increasing the performances of the typical algorithms within the signal and image processing fields thanks to the wealth of information it provides. However, there is no generic strategy to exploit this multimodality, as it greatly depends on the nature of the recorded signals as well as the objective to reach. In the remote sensing field in particular, multimodal data are a common phenomenon due to the multiplicity of imaging sensors. Again, the lack of generic method to make the most of this multimodality lead to the design of multimodal algorithms which are very specific with respect to the task and/or the nature of the handled multimodality. A unified framework able to handle equally all types of multimodalities would surely benefit the remote sensing community a lot.

On the other hand, hierarchical representations have proved to be a valuable tool when it comes to hierarchically decompose images. Such tools have shown to be of use for several typical image processing applications such as denoising, segmentation, filtering, and so on. The strength of hierarchical representations is that they act as an image decomposition tool regardless of the further application. The design of methods to make hierarchical representations handling multimodality could provide some suitable tools for many applications involving multimodal images. Therefore, the following of this manuscript investigates how multimodal information can be integrated into the construction and processing of hierarchical representations, to improve typical image processing applications.

In chapter 2, we investigate the use of spectral-spatial multimodality for segmentation purposes. More precisely, we focus on hyperspectral images, which can be seen as a set of gray-scale images depicting different spectral characteristics of the scene. The pursued objective in chapter 2 is the fusion of a typical spatial information based application, namely image segmentation, with a spectral information based application being spectral unmixing. Contrarily to most state-of-the-art methods that first perform the spectral unmixing and then integrate the spatial information as a regularization step, we proceed the other way around, seeking a partition of the space that is optimal in order to further perform spectral unmixing. We propose in particular a method based on the minimization of a suited energy function over the set of all cuts of a hierarchy of partitions in order to obtain this optimal partition.

In chapter 3, we handle sequences of hyperspectral images, thus introducing the temporal multimodality. As it notably brings some information related to motion, *i.e.*, what and how is changing from a frame to the other, a typical application linked to this temporal multimodality is the tracking of some object along the various frames of the sequence. In particular, we design a methodology to perform object tracking, based on the hierarchical decomposition of the sequence. While this has already been studied in the context of traditional color video sequences, the scarceness of available hyperspectral sequences (added to all other difficulties related to hyperspectral imaging) makes it extra-challenging to design efficient and generic tools. We study the scenario of chemical gas plume tracking, which is a particular application where all spectral, spatial and temporal information are crucial.

In chapter 4, we focus on the sensorial multimodality, namely when several images of a same scene are acquired with different imaging sensors. In that case, each modality features some particular information about the scene, and the combination of these should benefit image segmentation by helping the design of more accurate regions. However, processing such multimodal images raises the question on the fusion of several hierarchical decompositions. This question is answered by the introduction of braids of partitions, which generalize hierarchies of partitions. Relying on an energy minimization procedure, we propose a full methodological framework based on this notion of braid of partitions to perform the segmentation of such multimodal images.

Chapter 2

Spectral-Spatial multimodality

Contents

In this chapter, we turn our attention to the spectral-spatial multimodality. In particular, working with hyperspectral images (HSI [1]), we aim at fusing both the spectral and spatial information contained in such images in order to output a partition that is optimal with respect to the spectral unmixing operation. The generation of this optimal partition is done through the construction of a BPT representation of the HSI and an appropriate pruning of it by means of the minimization of a suited energy function. The organization of this chapter is as follows: in section 2.1, we introduce hyperspectral images as particular instances of spectral-spatial multimodality and the associated applications that benefit from this multimodality.

1. We shall emphasize here that the acronym HSI will stand for hyperspectral imagery, and not for the Hue, Saturation, Intensity color space, as it could be encountered in computer vision.

In section 2.2 and section 2.3 we recall the basics of spectral unmixing and some notions related to segmentation by minimization of an energy function (in particular the work of Guigues [87] which focuses on the minimization of such energy function over hierarchies of partitions), respectively. Section 2.4 presents the proposed methodology, which aims at combining the notions of hierarchical energy minimization and spectral unmixing in order to produce from the BPT representation of the HSI an optimal partition with respect to spectral unmixing. In particular, we propose a new way to construct the BPT representation through the introduction of novel region models and associated merging criteria, as well as new energy functions related to spectral unmixing. Conducted experiments are presented in section 2.5, where we apply the proposed methodology on two state-of-the-art hyperspectral data sets and perform comparison against classical strategies to build and to prune the BPT representation. Results are displayed in section 2.6, while section 2.7 draws some conclusions and future research avenues.

We would also like to mention that materials presented in this chapter were presented in our article [217], which is the fruit of a collaboration between several researchers [2]. Therefore, we will emphasis in particular the contributions of [217] which were made by the author of the present manuscript. However, for the sake of clarity and readability, we will also present the other contributions.

2.1 Hyperspectral spectral-spatial multimodality

2.1.1 Introduction

As mentioned in section 1.1.2, a hyperspectral image (HSI) is a collection of single band, gray-scale images, acquired simultaneously over narrow and contiguous wavelengths of the electromagnetic spectrum. From this acquisition procedure results a data cube where to each pixel location is associated a discrete spectrum related to the way the incident light has interacted with the region of the scene at this location. This interaction can be interpreted either in terms of the amount of light reflected by the scene (which is the dominant phenomenon when working with wavelengths in the visible and near infrared domains), one then talks of reflectance spectrum, or the amount of energy irradiated by the scene (when working in the middle and long wave infrared domains), and one talks of emissivity in that case.

Each physical material is characterized by its proper way to interact with light. Due to this, it is possible to establish from the spectrum depicted by each pixel in a HSI which are the materials constituting this spectrum, and thus to identify in a more general way the constituents of the scene. Given this capacity of recognizing the physical components present in the image, hyperspectral imagery has found numerous applications in several fields such as medical imaging [38, 125] (where it can be used for tumor extraction and identification,

2. This work was done in collaboration with Dr. Veganzones, Dr. Dalla Mura and Dr. Chanussot from the GIPSA-lab, Grenoble Institute of Technology, France, and with Dr. Plaza from the University of Extremadura, Cáceres, Spain.

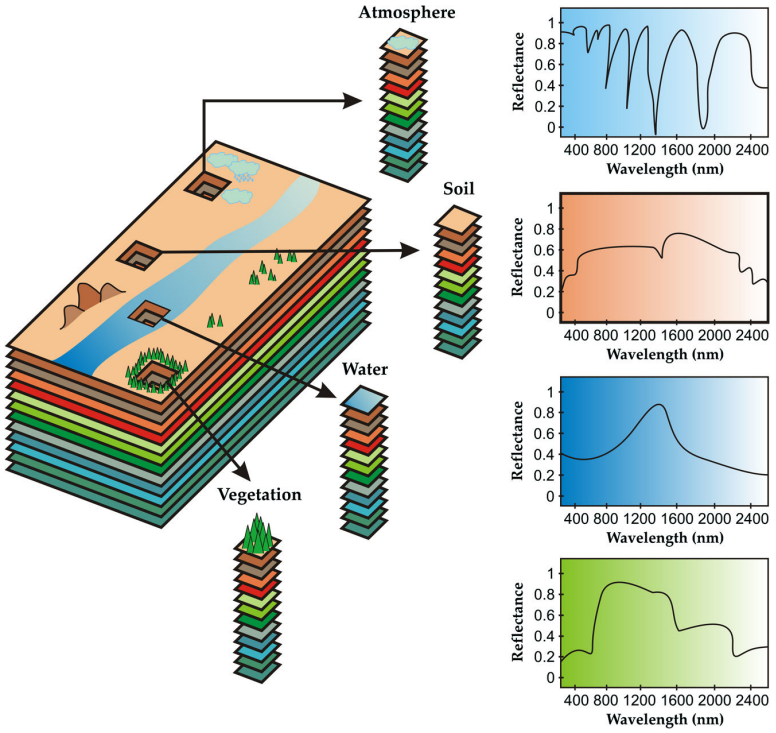


Figure 2.1: Illustration of a hyperspectral image (image extracted from [23]).

assessing tissue perfusion and its pathological conditions, helping for accurate surgical decisions or evaluating the health of dental structures for example), food quality inspection [69,124] (with applications in meat tenderness prediction as well as the detection of microbial spoilage for instance), geological [212] and hydrological [83] sciences and Earth and planetary observation by remote sensing [24, 82]. An illustration of a remotely sensed hyperspectral image of the Earth surface is depicted by figure 2.1. It can be seen how the three spectra corresponding to soil, water and vegetation differ from each other, thus acting as a signature for their corresponding material.

Typical hyperspectral sensors have a spectral resolution often comprised between 10 nm and 20 nm, meaning that each spectral channel of the image is concerned with the measurement of light in a very restrained portion of the spectral channel, whose bandwidth is at most 20 nm. However, the price to pay for a fine spectral resolution is a coarser spatial resolution. In spite of the technological advances made in the design of hyperspectral sensor, the typical spatial resolution is still at best in the order of a few meters (around 5 m for the AVIRIS and HyMap sensors for instance), while panchromatic and multi-spectral images now enjoy centimetric resolutions.

A first approach to enhance the spatial resolution of a low spatial/high spectral resolution hyperspectral image is the use of a complementary high spatial/low spectral resolution panchromatic or multi-spectral image of the same scene. This spectral-spatial multimodality, known as super-resolution or hyperspectral pansharpening, aims at generating from the two

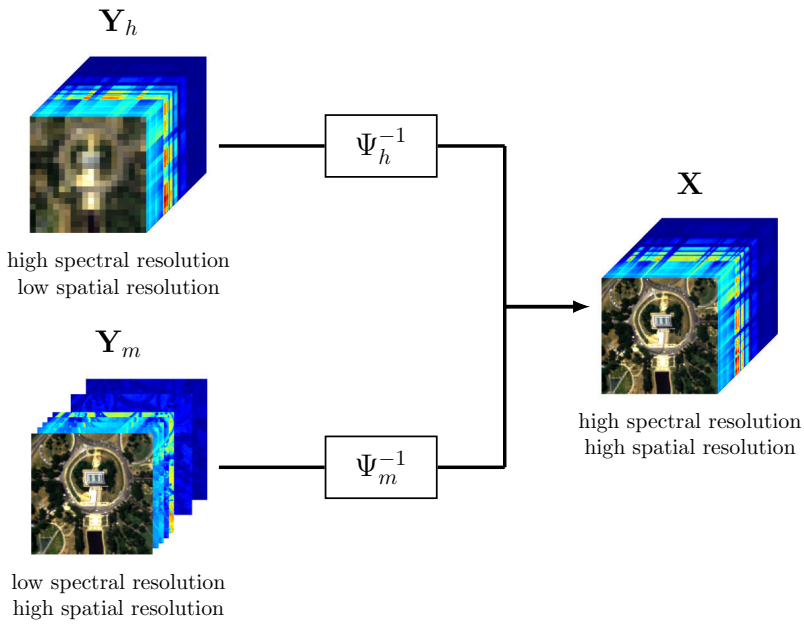


Figure 2.2: Illustration of a super-resolution algorithm (image borrowed from [232])

complementary images a high spectral/high spatial resolution image. The literature features plenty of algorithms devoted to the pansharpening of multispectral images, which can often be classified among component substitution, multiresolution analysis, Bayesian-based and variational-based methods (reviews of such methods can be found in [8, 191]). However, the extension of these methods to hyperspectral images is still more complex as phenomena such as the difference between the spectral ranges of the low and high spectral resolution images have to be taken into account. The line of conduct that is most of the time followed is to consider that the hyperspectral image $\mathbf{Y}_h$ and the high spatial resolution (be it multi-spectral or panchromatic) image $\mathbf{Y}_m$ are obtained from the super-resolution image $\mathbf{X}$ by some unknown transformations Ψ_h and Ψ_m:

$$\begin{aligned} \mathbf{Y}_h &= \Psi_h(\mathbf{X}) \\ \mathbf{Y}_m &= \Psi_m(\mathbf{X}) \end{aligned} \tag{2.1}$$

The goal of the super-resolution algorithm is then to estimate and invert those transformations in order to retrieve back the super-resolution image, as illustrated by figure 2.2. An extensive review of hyperspectral super-resolution methods is presented in [122].

However, the complementary high spatial/low spectral resolution image is often not always available, and one has to cope only with the spatial information contained within the HSI. However, considering a hyperspectral image as a source of multimodality does not contradict the definition of multimodal data given by definition 1.2. As a matter of fact, a HSI, being a collection of gray-scale images depicting the same scene at different wavelength positions, can be viewed as a source of both spectral and spatial information, compared to the case of a single band image (such as a panchromatic image, for instance). In any case, the joint consideration

of spectral and spatial information has shown to improve typical hyperspectral applications that initially relied either on spectral or spatial information only [3]. That is notably the case for hyperspectral classification, spectral unmixing and segmentation, which are reviewed below.

2.1.2 Examples of spectral-spatial multimodality for hyperspectral applications

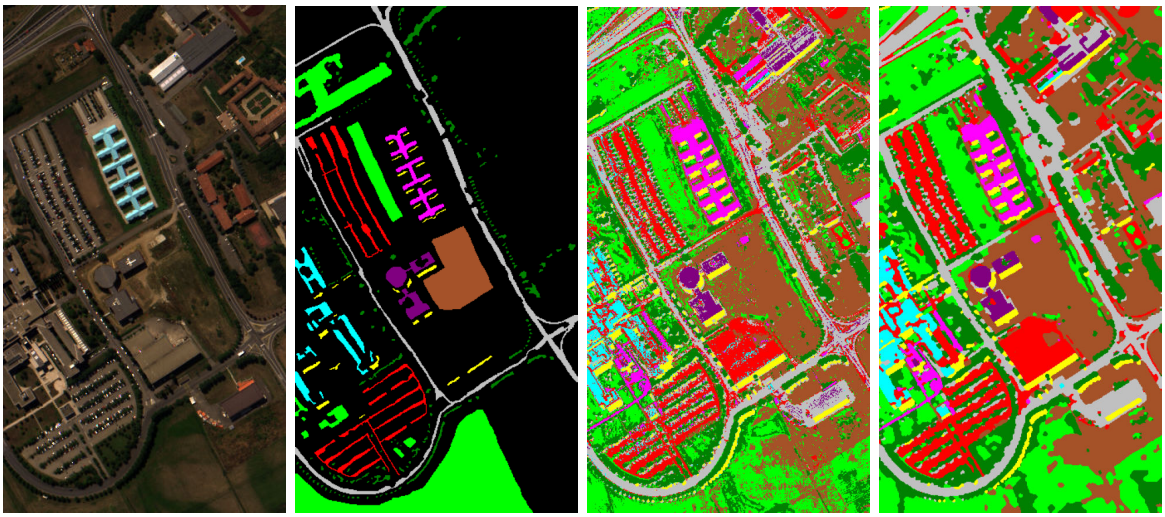
2.1.2.1 Hyperspectral classification

The classification task aims, given an image, at assigning a label to each pixel, such that pixels sharing the same label (which define a class) have some common properties. Note that the resulting classification map forms a particular partition of the image, where the regions may however not all be connected. The classification procedure can be unsupervised, semi-supervised or supervised. In the first case, the number and identity of classes are unknown *a priori* and have to be found within the algorithm itself or manually estimated. Well-known examples of unsupervised classification methods include clustering-based techniques such as k-means [121] or mean shift clustering [52]. When applied to hyperspectral images, these methods suffer from the very high dimensionality of the data. Supervised classification methods are known to outperform the typical results of unsupervised classification, but at the expense of the need of *a priori* known labeled samples. Those are divided into training samples, on which is first performed a learning step, and validation samples, which are used to assess the accuracy of the classification. Among supervised classification methods, one can notably cite artificial neural networks [65] as well as (kernel-based) support vector machines (SVMs) [54, 179]. Semi-supervised methods stand in-between supervised and unsupervised methods, as they rely both on some labeled and unlabeled samples (see [236] for a review).

Overall, when applied on hyperspectral images, classification methods suffer from the so-called "salt and pepper" effect. Each pixel is classified based only on its spectral characteristics, and the resulting classification map may show some spatial inconsistencies, such as pixels belonging to a given class isolated within another class. To overcome this issue, spatial information can be considered as a means to regularize the output of a pixel-wise classification map. Such spectral-spatial classification algorithms rely on the intuition that neighboring pixels are more likely to belong to the same class. Then, given a pixel-wise classification map, there are several strategies to make the most of the spatial information [73].

- The contextual information is embedded within a probabilistic framework, such as the one presented in [189] using a Markov Random Field (MRF) regularization. The final classification map, which is interpreted as the maximum a posteriori (MAP) estimate of the "true" (unknown) classification map is obtained by iteratively relabeling some pixels of the pixel-wise classification map by considering class dependencies between adjacent pixels.
- The contextual information is considered as a classification feature, where a common

3. Although the joint use of spectral and spatial information rather appears as a data fusion problem, we will term this as *spectral-spatial multimodality* for the sake of convenience.



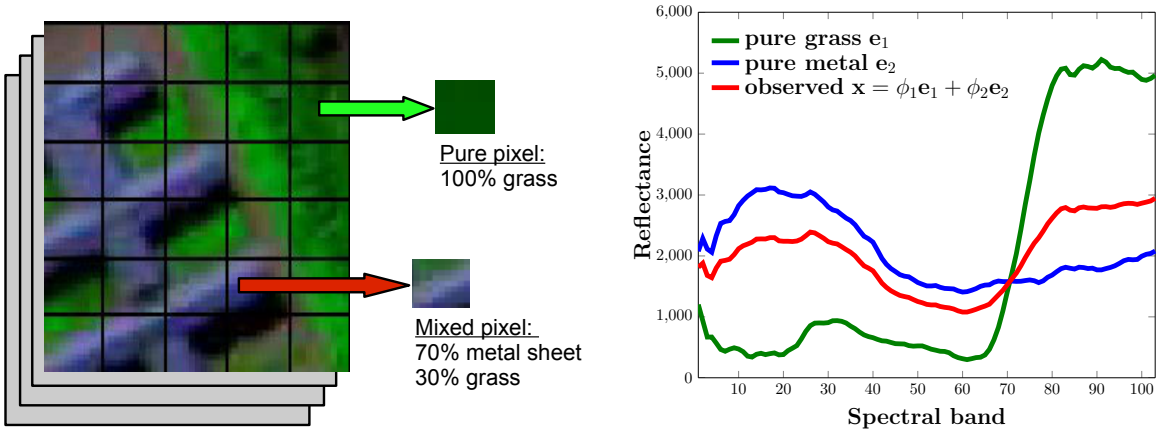
(a) (b) (c) (d)

Figure 2.3: Example of spectral-spatial hyperspectral classification: (a) RGB composition of the Pavia university hyperspectral data set, (b) labeled data (where each color defines a class), (c) example of pixel-wise SVM classification and (d) example of spectral-spatial classification (MAP-MRF method presented in [189]).

approach is to add spatial features in the classification process. The spatial features aim at embedding the spatial relations (contextual relations, geometrical features, etc.) of the objects in the scene. As an example, the use of morphological filters in a multi-scale setting leads to the definition of morphological profiles [19], which act as an adaptive neighborhood for each pixel. Then, two pixels are more likely to belong to the same class if their morphological profiles are similar. Therefore, the classification is performed using both spectral and spatial features.

– The contextual information is represented through a segmentation map. If two pixels belong to the same regions, then they likely belong to the same class. While this strategy is bound to the derivation of a good segmentation map, it was shown in [188] that the hyperspectral watershed performs consistently. First, a multidimensional gradient of the hyperspectral image is computed, and a watershed transformation is applied onto the gradient map. Provided a pixel-wise classification, all pixels of a region of the segmentation map are finally assigned to the most frequent class within the region (known as majority voting).

Spectral-spatial classification strategies finally output a classification map of the image, where each class looks more spatially homogeneous than in its pixel-wise counterpart. An example of such spectral-spatial classification improvement is displayed in figure 2.3. One can notably see in figure 2.3c the main drawback of pixel-wise classification where the spatial distribution of classes suffers from "salt and pepper" inconsistencies. This effect is strongly mitigated when incorporating spatial information, as it can be seen in figure 2.3d which displays the result of the spectral-spatial MAP-MRF-SVM approach of [189].



(a) (b)

Figure 2.4: (a) Illustration of a pure and a mixed pixel in a hyperspectral image (image borrowed from [219]), along with (b) the observed spectra in both cases.

2.1.2.2 Spectral unmixing

Another classical hyperspectral application that benefits from the spectral-spatial multimodality is spectral unmixing. The goal of spectral unmixing is to retrieve, for each pixel spectrum, which are the pure constituents, called *endmembers* (which correspond to macroscopic elements such as soil, vegetation, grass, concrete and so on) present in the spectrum and in which proportion (the *fractional abundances*). Given a HSI, the output of the spectral unmixing operation is the set of endmembers and their resulting abundance maps. The low spatial resolution of hyperspectral images is actually one of the motivations to perform spectral unmixing, as it is likely that several pure constituents are "mixed" within each pixel site and thus add their contribution to the resulting pixel spectrum, as illustrated by figure 2.4. By giving access to sub-pixelar information, spectral unmixing can also be viewed as a sort of super-resolution method.

The unmixing is commonly done over the whole set of pixels without any prior information related to the spatial distribution of the endmembers across the image. However, it can be assumed in a similar fashion as hyperspectral classification that neighboring pixels are likely to be made of the same endmembers in comparable proportions. Thus, introducing some spatial information within the unmixing process should lead to more consistent results. This idea has already been exploited in the literature in several works, with three main strategies standing out:

- The spatial information is integrated as a pre-processing step and combined with the derivation of pixel purity indices, in order to guide the search for endmembers in regions which are spatially homogeneous and spectrally pure. This is the case for instance in [134, 135].
- The spatial information is incorporated within the endmember selection process, as

in [158] which uses mathematical morphology operators to finds the purest pixels (most likely to be endmembers) in a given neighborhood, or in [168] where spatial characteristics are used to increase the spectral contrast between spectrally similar, but spatially independent endmembers, thus improving the potential of finding these endmembers.
- The spatial information is taken into account to smooth the abundance maps, by using some total variation regularization [94] or some MRF formulation [68] to model spatial correlations between neighboring pixels.

2.1.2.3 Hyperspectral hierarchical segmentation

Image segmentation process aims at dividing an image into regions fulfilling some given criterion. Often when working with hyperspectral images, one is interested in spectrally homogeneous regions, as it is commonly assumed that all pixels constituting a semantic object of interest should feature similar spectral properties. Most segmentation algorithms that have been proposed in the scope of hyperspectral image segmentation are by nature hierarchical as they rely on some region merging procedure, based on the evaluation of spectral similarities between regions. Thus, the hierarchical segmentation of a hyperspectral image is in essence a spectral-spatial processing, as it aims at decomposing the hyperspectral image in a set of nested regions (the spatial side of the spectral-spatial processing) which are spectrally coherent (the spectral part). We list below some notable hierarchical segmentation algorithms that have been proposed in the literature for hyperspectral image segmentation:
- The first proposed hierarchical segmentation method adapted to hyperspectral imagery was the Extraction and Classification of Homogeneous Objects (ECHO) algorithm [99]. ECHO implements a region merging procedure, where the decision whether to merge two regions or not is taken according to a likelihood test evaluating if two regions are homogeneous or not. It suffers from the common downside of every statistical test being the setting of a (false alarm) threshold which impacts the performances of the produced sequence of partitions. Moreover, it relies also on the computation of the inverse of covariance matrices, which can be problematic because these matrices are often badly conditioned when dealing with hyperspectral data.
- The Fractal Net Evolution Approach (abbreviated FNEA) proposed by [12] also carries out region merging. The fusion procedure in the FNEA algorithm minimizes at every step the growth in heterogeneity in a heuristic process. Provided a spectral homogeneity measure between two regions (such as the Euclidean distance between the mean spectra of two regions in the original paper [12]), a virtual merge between those two regions is first evaluated in order to measure and compare the homogeneity of the virtual region against the ones of its constituents. The final merging occurring at the current step is the one which minimizes this loss in homogeneity.
- The Hierachical SEGmentation (HSEG) method, proposed in [192, 193] and based on the well-known Hierarchical Set-Wise Optimization (HSWO) procedure [18]. In the latter, each iteration involves the search for the two adjacent regions that have the lowest pairwise distance. All pairs of regions achieving this distance are then merged. The

HSEG algorithm is founded on the same idea, expect that the adjacency constraint for regions is partially relaxed. Indeed, a user-chosen proportion of non-adjacent regions can also be merged at each iteration, provided that their distance is less than the minimal distance among all pairs of adjacent regions. For that reason, the HSEG algorithm can be viewed as sequentially alternating between a region growing step and a spectral clustering step. Due to its huge computational load, induced by the important number of pairwise distances that must be evaluated the HSEG algorithm was further extended to the Recursive HSEG (RHSEG) algorithm [194]. This latter, based on a divide-and-conquer approximation of the HSEG, allows for parallel implementation and computational acceleration.

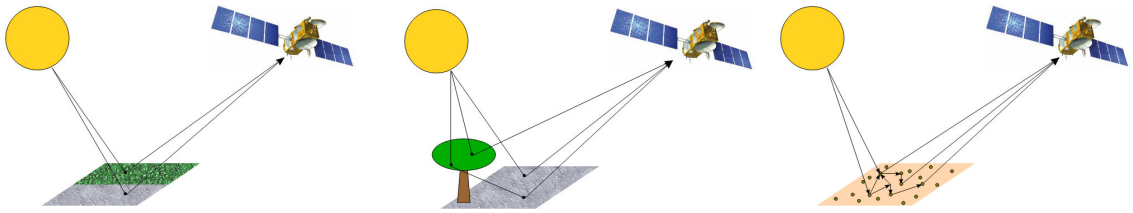
– The adaptation of the BPT representation to hyperspectral images, as presented in [207], introduced in chapter 1. The construction of the BPT strongly resembles the HSWO procedure, the only difference being that each merging iteration feature the merging of only two regions in the case of the BPT even if several pairs of regions have the same minimum pairwise lowest distance. The major difference between the work of Valero [204, 207] and all previously cited hierarchical segmentation methods is in the further processing. In the latter case, it is assumed that the "optimal" segmentation can be found directly in the stack of partitions created during the region merging process [190]. However, this is rarely true, especially when objects of interest can be found at different levels of the hierarchy. To that extend, new tree-based processing techniques were introduced in the work of Valero to make the most of the hierarchical decomposition of the HSI induced by its BPT representation.

2.1.3 Objective of this chapter

The goal of this chapter is to propose a new way to incorporate spectral-spatial information in a BPT-based representation of a hyperspectral image. While all hierarchical segmentation works cited above handle the spectral information through the pixel spectra, we propose to go one step beyond by performing spectral unmixing over each region and to handle this region through its proper endmembers and associated fractional abundances. This is in itself a depart from the conventional spectral unmixing, performed over the whole image.

The pursued objective, by means of the BPT representation, is to obtain at the end of the day a segmentation map of the hyperspectral image which can be called *optimal* with respect to the spectral unmixing operation. This optimality, developed in the sequel, is reached by the definition of an energy function linked with spectral unmixing, which is subsequently minimized over the hierarchy. Therefore, spectral and spatial information are used in a synergistic way at different stages of the proposed methodology:

1. During the construction of the BPT, spectral and spatial information are used for the definition of suitable region models and merging criteria.
2. During the processing of the BPT, where an energy function integrating spectral and spatial considerations is minimized over the previously constructed BPT representation.
3. The output of the proposed method: a segmentation map (encoding spatial information



(a) (b) (c)

Figure 2.5: Different interaction models between the incident light and the surface. The interactions can be modelled as linear (a) when each ray of light bounces only one, or non-linear (b)-(c) because of multiple or intimate rebounds.

by nature) which is optimal with respect to the unmixing operation (which is by essence a spectral processing).

2.2 Spectral unmixing

When a hyperspectral sensor, be it airborne or spaceborne, collects an image of the Earth, it actually records the amount of light which bounced off the surface and reached the sensor. Each ray of light interacted with the elements present on the ground during its rebound. The goal of spectral unmixing is, given what has reached the sensor, to identify which were the elements on the surface and how the light interacted with them.

These interactions can be modeled from the physics as non-linear for several reasons, such as the topography of the ground which can lead to multiple rebounds (illustrated by figure 2.5b) or the consistency of the material which may generate some intimate mixing phenomena (depicted by figure 2.5c). However, due to the complexity of these models, non-linear effects are often neglected and approximated by a linear mixing model (LMM) instead. This latter assumes that each ray of light bounces only on one element on the ground so the optical signal that is received by the sensor for each pixel site is the mean of all the interaction that happened within this site [98] (see figure 2.5a). Despite its relative simplicity, most of the unmixing methods in the literature are based on the LMM [23] as it allows simple geometric interpretations. Given a HSI, a classical spectral unmixing algorithm outputs the set of spectral signatures of the main constituents of the scene, called *endmembers*, and their corresponding *fractional abundances* that depicts the spatial distribution of these endmembers within the scene.

2.2.1 Linear Mixing Model (LMM)

The LMM states that a hyperspectral sample is formed by a linear combination of the spectral signatures of pure materials present in the sample (endmembers), plus some additive

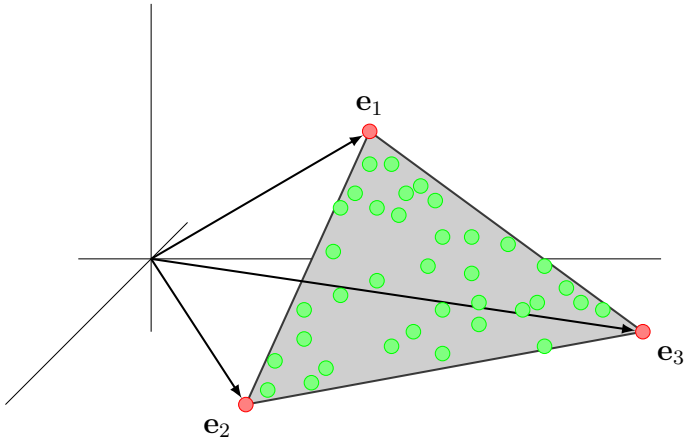


Figure 2.6: Geometrical interpretation of the LMM combined with the ANC and ASC. The simplex (in gray) is formed by the three endmembers (in red) $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$, and all image pixels (in green) lie within it.

noise. Let $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_m]$ be the pure endmember signatures (normally corresponding to macroscopic objects in the scene, such as water, soil, vegetation,...) where each $\mathbf{e}_i \in \mathbb{R}^N$ is a N-dimensional vector. Then, the hyperspectral signature $\mathbf{x}$ at each pixel in the image is defined by the expression:

$$\mathbf{x} = \mathbf{s} + \boldsymbol{\eta} = \sum_{i=1}^{m} \phi_i \mathbf{e}_i + \boldsymbol{\eta}, \tag{2.2}$$

where $\mathbf{x}$ is given by the sum of the pixel signal $\mathbf{s}$ and an independent additive noise component $\boldsymbol{\eta}$. $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_m]$ is the m-dimensional vector of fractional per-pixel abundances related to $\mathbf{x}$, which models the contribution in percentage of each endmember $\mathbf{e}_i$ in the signature $\mathbf{x}$. For physical reasons, it is subject to the Abundance Non-negative Constraint (ANC) and the Abundance Sum-to-one Constraint (ASC):

$$\phi_i \geq 0 \ \forall i = 1, \ldots, m \quad \text{(ANC)} \ , \tag{2.3}$$

$$\sum_{i=1}^{m} \phi_i = 1 \quad \text{(ASC)} \ . \tag{2.4}$$

The LMM combined with the ANC and ASC can be interpreted from a geometrical point of view: the m endmembers of the image form a $(m-1)-$simplex whose vertices are the endmembers. All pixels of the image, which can be written as a linear combination of the endmembers weighted by the fractional abundances, then lie inside the simplex, as illustrated by figure 2.6. This interpretation notably paves the way to several geometrical methods for the endmembers identification, which are reviewed in [23].

Representing the HSI as a matrix $\mathbf{X} \in \mathbb{R}^{N \times N_{\text{pix}}}$, it is possible to extend equation (2.2) to the whole image as

$$\mathbf{X} = \mathbf{E}\boldsymbol{\Phi} + \boldsymbol{\eta} \ , \tag{2.5}$$

where $\mathbf{E} \in \mathbb{R}^{N \times m}$ is the matrix whose columns correspond to the m endmember signatures, $\mathbf{\Phi} \in \mathbb{R}^{m \times N_{\text{pix}}}$ is the matrix of fractional abundances and $\boldsymbol{\eta}$ is an independent additive noise. This formulation allows the use of matrix factorization methods to infer matrices $\mathbf{E}$ and $\mathbf{\Phi}$ by solving the following problem:

$$\min_{\mathbf{E},\mathbf{\Phi}} \|\mathbf{X} - \mathbf{E}\mathbf{\Phi}\|_{\dagger}^{2} \quad \text{such that } \mathbf{\Phi} \succeq 0 \text{ (ANC)}, \mathbb{1}_{m}^{T}\mathbf{\Phi} = \mathbb{1}_{N_{\text{pix}}} \text{ (ASC)} , \tag{2.6}$$

where $\|\cdot\|_{\dagger}$ is often formulated as the Euclidean or Frobenius norms [95, 162].

2.2.2 Endmember induction and abundance estimation

Most of the times, the spectral signatures of the materials are unknown, and the set of endmembers must be built by either selecting spectral signatures from a spectral library, or by automatically inducing them from the image itself. Both can be performed manually or in an automatic way. In order to automatically induce the set of endmembers from the image, the use of some endmember induction algorithm (EIA) is required. The hyperspectral literature features plenty of such algorithms. Some reviews on the topic can be found in [23, 98, 214].

Once the set of endmembers, $\hat{\mathbf{E}}$, has been induced, their corresponding per-pixel abundances, $\hat{\mathbf{\Phi}}$, can be estimated by approximating a solution to an over-determined linear system by the Least Squares method [110]. The Fully-Constrained Least Squares Unmixing (FCSLU) method [89] solves the over-determined linear system subject to ANC and ASC constraints.

The quality of the unmixing, $\hat{\mathbf{E}}$ and $\hat{\mathbf{\Phi}}$, at a given pixel x can be measured by the Root Mean Squared Error (RMSE) of the original hyperspectral signature with respect to the reconstructed one, $\hat{\mathbf{x}} = \sum_{i=1}^{m} \hat{\phi}_i \hat{\mathbf{e}}_i$:

$$\epsilon(\mathbf{x}, \hat{\mathbf{x}}) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} \left(x\left(k\right) - \hat{x}\left(k\right)\right)^{2}}. \tag{2.7}$$

Computing the RMSE of each pixel x yields a reconstruction map where low values signify that the corresponding pixels have been well reconstructed by the set of induced endmembers. Contrarily, a high error value for a pixel means that its spectral signature is not well explained by a linear combination of the endmember spectra.

2.3 Energy minimization over hierarchies of partitions

Image segmentation is one of the most investigated applications in image processing [4]. The main reason of this popularity is actually due to the complexity of such operation. As a matter of fact, image segmentation is an ill-posed problem: a given image can be segmented in a variety of partitions, and the intended segmentation result depends on the pursued

4. Typing "image segmentation" in Google scholar and restricting to the 2010-2015 period yields around 214 000 results.

application. Roughly speaking, image segmentation algorithms can be classified into four different categories:

- Region-based methods. Those aims at producing the segmentation by focusing on regions, *i.e.*, sets of pixels, such that they all fulfill some predefined criteria (such as shape, homogeneity, texture, and so on). Hierarchical segmentation methods notably belong to this class, as the construction of a hierarchy is conducted through the aggregation (or splitting) of a set of regions of the image.
- Edge-based methods, which can be seen as the dual of region-based methods. Rather than focusing on the pixels composing the region, they are based on the properties of the transitions between regions. The final segmentation map is defined by its region boundaries instead of the regions themselves. All methods based on the detection of edges belong to this category.
- Statistical-based methods. This group of algorithms produces a segmentation by exploiting the inherent statistics of the image. For instance, a simple thresholding can be seen as a statistical-based segmentation as the threshold is often set according to statistical distribution of the pixel values in the image. Clustering algorithms, such as the mean shift clustering, are also particular instances of statistical-based methods.
- The remaining methods that do not fit within a previous category, mostly because they exploit several of the previous aspects. For instance, segmentation methods relying on a graph setting often assign to each edge a value reflecting the dissimilarity between the vertices connected by this edge. The segmentation is then derived by producing sets of connected vertices such that their dissimilarity is low (which is a region-based approach) and such that the dissimilarity between each vertex in the set and a connected vertex outside the set is high (which can be viewed as an edge-based idea).

2.3.1 Segmentation by energy minimization

As image segmentation is application-dependent, one often tries to find the "best" segmentation of an image for a given task. This notion of optimality often relies on the definition of an *energy* function (also called *objective* function, or *cost* function, given the domain of application), which embeds in its expression the properties that should be featured by the optimal segmentation. By reflecting how good or bad is a given segmentation with respect to the application, it is then possible to define the optimal one as the minimizer of the energy function.

The main advantage of this segmentation by energy minimization framework is that it shifts the problem of identifying an optimal segmentation among the set of all possible ones, which is subjective, to the problem of properly defining an energy function whose minimizer is the sought segmentation, which is objective and can be formulated mathematically. This energy minimization framework has been extensively used in the literature, and one can notably cite:

Mumford-Shah functional: Representing an image as a function $\mathcal{I}_0 : E \to V$, Mumford and Shah proposed in their famous paper [142] to define the optimal segmentation of $\mathcal{I}_0$

as the minimizer of the following energy function:

$$\mathcal{E}^{MS}(\mathcal{I},\Gamma) = \int_E \|\mathcal{I}(x) - \mathcal{I}_0(x)\|^2 dx + \lambda \int_{E\setminus\Gamma} \|\nabla\mathcal{I}(x)\|^2 dx + \nu|\Gamma| \tag{2.8}$$

where $\mathcal{I}$ is a piece-wise smooth approximation of $\mathcal{I}_0$, Γ is a set of boundaries whose total length is $|\Gamma|$. The previously defined energy function is composed of three terms: the first one penalizes the misfit between the original image $\mathcal{I}_0$ and its approximation $\mathcal{I}$, the second one enforces smoothness for $\mathcal{I}$ and the last term promotes simplicity in the segmentation by regularizing the total length of boundaries. Minimizing equation (2.8) has been the concern of several studies ([47] uses a levet set approach for instance, while an approximation by finite differences is implemented in [46]), but it is known to be a non-convex NP hard problem. It has therefore been subsequently relaxed into the so-called piece-wise constant Mumford-Shah functional:

$$\mathcal{E}^{MS}(\pi) = \sum_{\mathcal{R}_i\in\pi} \left(\int_{\mathcal{R}_i} \|\mathcal{I}_0(x) - c_i\|^2 dx + \frac{\nu}{2}|\partial\mathcal{R}_i| \right) \tag{2.9}$$

where the space E is now partitioned into a set of connected components $\pi = \{\mathcal{R}_i\}$ and the approximation $\mathcal{I}$ of $\mathcal{I}_0$ is set to the constant value c_i over $\mathcal{R}_i$. For a given partition π, the values of c_i that actually minimize equation (2.9) are the mean values of $\mathcal{I}$ over $\mathcal{R}_i$, denoted $\mu_{\mathcal{I}_0}(\mathcal{R}_i)$. The finally obtained piece-wise constant Mumford-Shah expression

$$\mathcal{E}^{MS}(\pi) = \sum_{\mathcal{R}_i\in\pi} \left(\int_{\mathcal{R}_i} \|\mathcal{I}_0(x) - \mu_{\mathcal{I}_0}(\mathcal{R}_i)\|^2 dx + \frac{\nu}{2}|\partial\mathcal{R}_i| \right) \tag{2.10}$$

however remains difficult to minimize in practice as it is still non-convex when minimized without any further constraints with respect to π. However, we will see in the following that, when π is defined as a cut of a hierarchy, the minimizer of equation (2.10) can be found easily by a dynamic program.

Graph cut: In their well-known paper [27], Boykov, Veksler and Zahib propose to view the image segmentation as a labeling problem, *i.e.*, as a function $\mathcal{L}$ that assigns to each pixel $x \in E$ a label l_x in some given set of labels. Attaching a given energy to the labeling function, the optimal image segmentation is defined as the labeling with minimal energy. In particular, they consider a broad range of energy functions which can be written as

$$\begin{aligned} \mathcal{E}(\mathcal{L}) &= \mathcal{E}_{smooth}(\mathcal{L}) + \mathcal{E}_{data}(\mathcal{L}) \\ &= \sum_{(x,y)\in\mathcal{N}} V(l_x, l_y) + \sum_{x\in E} D(l_x) \end{aligned} \tag{2.11}$$

where $\mathcal{N}$ is the set of interacting pairs of pixels (not necessarily restricted to neighboring pixels) where V is some penalty function over the labels l_x and l_y of two interacting pixels x and y, and D is some data fitting term which measures how well the label l_x fits pixel x given the observed data. The proposed minimization is conducted with a graph-cut approach. Two operations, namely swap and expansion moves, are introduced. Both operations allow to reach some local minima, and it is shown that expansion moves can bring the local minima to an energy which can be at least twice the energy of the global minimum.

MRF: Markov Random Fields have also been widely studied to achieve image segmentation [113]. The underlying idea is similar to that of the previously presented graph cut approaches, namely to tackle the image segmentation as an labeling problem. A probability measure is associated with each possible labeling $\mathcal{L} \in \mathfrak{L}$, and the optimal one $\mathcal{L}^\star$ is the one maximizing the probability of a labeling given the observed image $\mathcal{I}$, $p(\mathcal{L}|\mathcal{I})$:

$$\begin{aligned} \mathcal{L}^\star &= \underset{\mathcal{L}\in\mathfrak{L}}{\operatorname{argmax}}\, p(\mathcal{L}|\mathcal{I}) \\ &= \underset{\mathcal{L}\in\mathfrak{L}}{\operatorname{argmax}}\, p(\mathcal{I}|\mathcal{L})p(\mathcal{L}) \end{aligned} \tag{2.12}$$

which is the maximum a posteriori (MAP) estimate of $\mathcal{L}$. Imposing $\mathcal{L}$ to be a MRF allows to formulate $p(\mathcal{L})$ as a Gibbs distribution, and to transform the MAP estimation (2.12) to the minimization of some energy function $U(\mathcal{L},\mathcal{I})$ which is often split in two terms, one accounting for region homogeneity and the other being a regularizer. The final minimization is often conducted by a simulated annealing approach [108], where the theoretical convergence to the global optimal is ensured in certain cases, but at a really slow rate.

To summarize, energy minimization is a convenient and widely used framework to achieve image segmentation as the specificities which must be achieved by the segmentation can be embedded in the energy definition. However, finding the minimizer of the energy function is not straightforward, either because the minimization problem is non-convex and the convergence to a global optimum is not guaranteed, or because this global optimum cannot be reached in an acceptable computational time. These limitations find their source in the structure of the space of all possible partitions Π_E of the set E:

- Its cardinality is gigantic: the number of partitions of a set E constituted of $|E|$ elements is given by the Bell number $B_{|E|}$. For instance, a 5×5 image possesses $B_{25} = 4.6 \times 10^{18}$ different partitions. This number drops if regions are constrained to be connected, but it remains in practice highly unrealistic to investigate all possible combinations.
- It is "unstructured": even if Π_E is known to be a lattice when equipped with the refinement ordering, most pairs of partitions $\pi_1, \pi_2 \in \Pi_E$ are not comparable, thus making an "intelligent" browsing of the partitions challenging.

In the following, we are going to see that the use of hierarchies of partitions can be a solution to both previously raised limitations.

2.3.2 Hierarchical segmentation by energy minimization

As we saw, the main reason why conducting segmentation by energy minimization is challenging is due to the structure and size of the space of partitions Π_E. A possible solution to alleviate this issue is no longer to conduct the search for an optimal partition on Π_E, but rather on the space of cuts $\Pi_E(H)$ of a hierarchy of partitions H. Indeed, the latter idea features two main advantages with respect to the classical framework:

- The set of possible cuts is strongly constrained by the structure of the hierarchy. Its cardinality is drastically reduced with the respect to the one of Π_E, even if it is impossible

in practice to evaluate as it depends on the architecture of H (such as the number of levels in the hierarchy, the average number of children per node, and so on).
- There is some underlying relationship (which will be called *h-equivalence* in chapter 4) between all partitions of $\Pi_E(H)$: given two cuts π_1 and π_2 of $\Pi_E(H)$, each region of π_1 is either disjoint or nested with all regions of π_2. This inclusion relationship, holding between all regions of the cuts composing $\Pi_E(H)$, should be exploited in order to find the optimal cut, if it exists, in a smart way.

Given some hierarchy of partitions H, the conditions which must be satisfied by the energy function $\mathcal{E}$ to ensure the existence of an optimal partition were first studied formally in the work of Guigues [86, 87], and later generalized in the work of Kiran [101, 103]. In the following of this section, we summarize the main results of the former, on which we will base ourselves to propose some new energy definitions [5].

2.3.2.1 Definitions and recalls

We shall start by recalling some definitions related to hierarchies. A hierarchy of partitions, H, constructed over a set E can be defined in two equivalent ways:
- As a sequence of partitions $\{\pi_i \in \Pi_E, i = 0, \dots, n\}$ which are ordered by refinement: $i \leq j \Rightarrow \pi_i \leq \pi_j$. π_0 is termed the *leaf* partitions, its regions are the *leaves* of H. Conversely, $\pi_n = \{E\}$ is the *root* of H.
- As a collection of regions $\{\mathcal{R} \subseteq E\}$ which includes $\{E\}$ but not $\emptyset$, and such that any two regions $\mathcal{R}_i$ and $\mathcal{R}_j$ are either disjoint ($\mathcal{R}_i \cap \mathcal{R}_i = \emptyset$) or nested ($\mathcal{R}_i \subseteq \mathcal{R}_j$ or $\mathcal{R}_j \subseteq \mathcal{R}_i$). In addition, if $\mathsf{C}(\mathcal{R})$ is the set of children of a region $\mathcal{R} \in H$, then $\mathcal{R} = \bigcup\{\mathcal{R}_c \in \mathsf{C}(\mathcal{R})\}$.

From each non-leaf region $\mathcal{R} \in H$, we can define a sub-hierarchy $H(\mathcal{R})$ rooted at $\mathcal{R}$.

A *cut* of H is a partition π of E whose all regions belong to H. From a graphical point of view, a cut can be seen as a path that intersects each branch of the tree-based representation of H at most once. The set of all cuts of a hierarchy H of a space E is denoted $\Pi_E(H)$. It is a sub-lattice of Π_E for the refinement ordering, meaning that the refinement supremum and refinement infimum of two cuts of H are also cuts of H. A cut of a sub-hierarchy $H(\mathcal{R})$ is called a *partial partition* of $\mathcal{R}$, and is denoted $\pi(\mathcal{R})$. As for the set of cuts of a hierarchy, the set of partial partitions of $\mathcal{R} \in H$ is denoted $\Pi_E(H(\mathcal{R}))$.

In the previous section 2.3.1, we introduced energy functions as a description of how good or bad a partition fits a given goal, postulating that the *optimal* partition is the one of minimal energy. As a matter of fact, energy functions are often considered to be real non-negative, with the intuition that the lower the energy, the better (or the "more stable") the corresponding partition commonly borrowed from physics. Therefore, a first mathematical definition of an energy function could be a mapping $\mathcal{E} : \Pi_E \to \mathbb{R}^+$ from the set of partitions of E to real non-negative numbers. However, in many cases, the energy function is evaluated over the regions composing the partition, which are then somehow assembled into the energy of the partition. This is for instance the case with the piece-wise constant Mumford-Shah energy

5. The novel energy functions we propose in this chapter are actually a particular case of those proposed in the work of Kiran [101], which were developed in parallel of our proposed work.

formulated by equations (2.9) and (2.10) where the energy of the partition is expressed as the sum of the energies of the regions composing the partition. Therefore, we consider the following general definition of energy functions:

Definition 2.1 (Energy function)
The definition of an energy function is based on two inner concepts:

1. *The definition of a regional energy, i.e., a function* $\mathcal{E}$*:* $\mathcal{P}(E) \to \mathbb{R}^+$ *that maps any region* $\mathcal{R} \subseteq E$ *to* $\mathbb{R}^+$*, where* $\mathcal{P}(E)$ *is the set of all subsets (i.e., possible regions) of* E*.*
2. *The definition of some rule* $\mathfrak{D}$ *to explicit the energy of a partition as some composition of the energies of its regions.*

The final energy of a partition $\pi \in \Pi_E$ *can be expressed as*

$$\mathcal{E}(\pi) = \bigop{\mathfrak{D}}_{\mathcal{R}_i \in \pi} \mathcal{E}(\mathcal{R}_i). \tag{2.13}$$

In this formalism, the composition rule $\mathfrak{D}$ can be arbitrary. The most common case is to express the energy of a partition as the sum of the energies of its regions (as in the Mumford-Shah energy for instance). However, we will see some other composition rules in the following section 2.4 and in chapter 4.

2.3.2.2 Optimal cut

The first question that was investigated by Guigues is the condition on $\mathcal{E}$ under which it is possible to guarantee the existence of an optimal cut

$$\pi^\star = \operatorname*{argmin}_{\pi \in \Pi_E(H)} \mathcal{E}(\pi) \tag{2.14}$$

and how to retrieve it in $\Pi_E(H)$. For that purpose, Guigues placed himself in the context of *separable energies*:

Definition 2.2 (Separable energy)
An energy $\mathcal{E}$ *is said to be separable if the energy of the partition* π *can be expressed as the sum of the energies of its regions:*

$$\mathcal{E} \text{ is separable} \Leftrightarrow \mathcal{E}(\pi) = \sum_{\mathcal{R} \in \pi} \mathcal{E}(\mathcal{R}) \ . \tag{2.15}$$

Note that the definition of a separable energy reduces to definition 2.1 with $\mathfrak{D} \equiv \sum$. Denoting $\pi^\star(\mathcal{R}) = \operatorname{argmin}_{\pi(\mathcal{R}) \in \Pi_E(H(\mathcal{R}))} \mathcal{E}(\pi(\mathcal{R}))$ the partial partition of $\mathcal{R}$ whose energy is minimal, and $\mathcal{E}^\star(\mathcal{R}) = \mathcal{E}(\pi^\star(\mathcal{R}))$ standing for this optimal energy, Guigues showed [86, pp. 141-142] that, for any separable energy $\mathcal{E}$, the following Bellman's dynamic program was

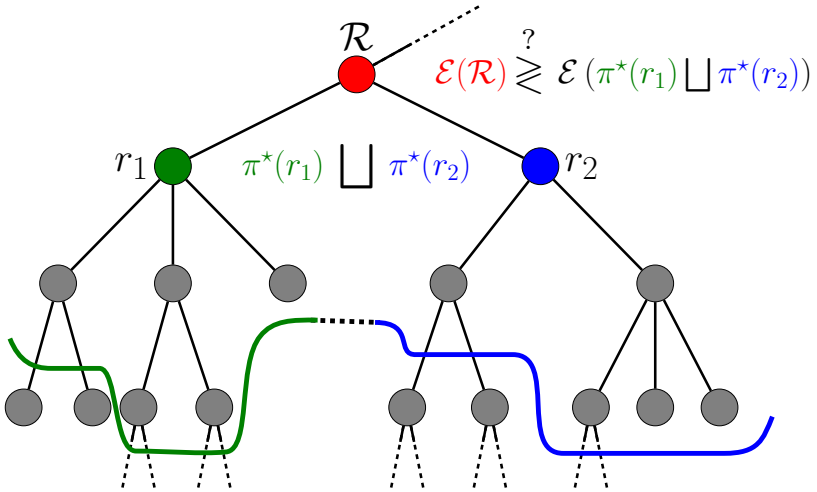


Figure 2.7: Illustration of a Bellman's dynamic program step to retrieve the optimal cut of a hierarchy.

holding for all region $\mathcal{R} \in H$:

$$\mathcal{E}^\star(\mathcal{R}) = \min\left\{\mathcal{E}(\mathcal{R}), \sum_{r \in \mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r)\right\} \tag{2.16}$$

$$\pi^\star(\mathcal{R}) = \begin{cases} \{\mathcal{R}\} & \text{if } \mathcal{E}(\mathcal{R}) \leq \displaystyle\sum_{r \in \mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r) \\ \displaystyle\bigsqcup_{r \in \mathsf{S}(\mathcal{R})} \pi^\star(r) & \text{otherwise} \end{cases} \tag{2.17}$$

with $\sqcup$ denoting disjoint union between regions (concatenation). Equations (2.16) and (2.17) means that the optimal energy of any region $\mathcal{R} \in H$ is given by comparing the proper energy $\mathcal{E}(\mathcal{R})$ of the region against the sum of the optimal energies of its children, and by picking the smallest of the two. The optimal cut of $\mathcal{R}$ is then given either by itself $\{\mathcal{R}\}$ or by the disjoint union of the optimal cuts of its children. This dynamic program procedure is illustrated by figure 2.7: looking for the optimal cut of $\mathcal{R}$ (in red), one has to compare its own energy $\mathcal{E}(\mathcal{R})$ against the energy of the union of the optimal cuts of its two children r_1 (in green) and r_2 (in blue), $\pi^\star(r_1) \sqcup \pi^\star(r_2)$. The energy being separable in the present case, this latter term is equal to $\mathcal{E}^\star(r_1) + \mathcal{E}^\star(r_2)$. Following, $\pi^\star(\mathcal{R})$ is either given by $\{\mathcal{R}\}$ or by $\pi^\star(\mathcal{S}_1) \sqcup \pi^\star(\mathcal{S}_2)$, depending on which has the lowest energy.

In practice, it is possible to obtain the optimal cut of the H by applying equations (2.16) and (2.17) over each region of the hierarchy, scanned in an ascending pass. The optimal cut $\pi^\star$ of H is given by the one of the root note. It is interesting to notice that the global optimal cut $\pi^\star$ is obtained by solving and concatenating a set of partial cuts which are locally optimal. As a matter of fact, each region $\mathcal{R}^\star \in \pi^\star$ has a lower energy than any of its partial partitions, and any of the partial partitions it is included in. This can be considered as a strong result knowing that the only condition required for the energy function $\mathcal{E}$ is separability.

The dynamic program procedure also illustrates that the optimal cut is obtained by taking advantage of the inclusion relationship holding on the regions of a hierarchy, thus emphasizing the benefit of conducting the energy minimization operation over such hierarchical structures instead of the unconstrained set of partitions Π_E.

The dynamic program methodology was actually used first for classification [28] purposes and wavelet bases constructions [51, 64] over quad-tree hierarchies, although it was not formulated as a dynamic program, in these works. It was also implemented both in [172] to solve the image rate/distortion problem and in [204] for hyperspectral segmentation, formulated as a Lagrangian minimization procedure.

2.3.2.3 Affine separable energies

The dynamic program procedure previously presented allows to find the optimal cut $\pi^\star$ of a hierarchy of partitions H given a separable energy function $\mathcal{E}$. Guigues then investigated the particular case of affine separable energies:

Definition 2.3 (Affine separable energy (ASE))
An energy $\mathcal{E}$ is said to be affine separable if it is separable and can be written as the sum of two terms weighted by some positive coefficient λ

$$\mathcal{E}_\lambda(\pi) = \sum_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R}) \tag{2.18}$$

In an affine separable energy, written in short $\mathcal{E}_\lambda = (\mathcal{E}_\phi, \mathcal{E}_\rho)$, the two terms $\mathcal{E}_\phi$ and $\mathcal{E}_\rho$ are competing to impose their own effect to the optimal partition, with a weight controlled by the parameter λ. An affine separable energy can be seen as a family of energies $\{\mathcal{E}_\lambda\}_{\lambda\in\mathbb{R}^+}$ parametrized by the coefficient λ. Therefore, it no longer generates a unique optimal cut $\pi^\star$, but rather a family of them $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ in turn indexed by the parameter λ. The behavior of $\pi^\star_\lambda$ with respect to λ is bound to the notion of sub-additivity, which then allows to formulate the multiscale minimal cuts theorem [86, pp. 161-162]:

Definition 2.4 (Sub-additive energy)
A separable energy $\mathcal{E}$ is sub-additive if for any two partitions π_1 and π_2 such that $\pi_1 \leq \pi_2$, then $\mathcal{E}(\pi_1) \geq \mathcal{E}(\pi_2)$. Equivalenty, for any two disjoint regions $\mathcal{R}_1$ and $\mathcal{R}_2$, $\mathcal{E}(\mathcal{R}_1 \cup \mathcal{R}_2) \leq \mathcal{E}(\mathcal{R}_1) + \mathcal{E}(\mathcal{R}_2)$.

Theorem 2.1 (Multiscale minimal cuts)
Let H be a hierarchy on a set E, and let $\mathcal{E}_\lambda = (\mathcal{E}_\phi, \mathcal{E}_\rho)$ be an affine separable energy. If $\mathcal{E}_\rho$ is sub-additive, then the family of optimal cuts $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ can be ordered by refinement, i.e.:

$$\forall\lambda_1, \lambda_2, 0 \leq \lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2} \tag{2.19}$$

The main consequence of this theorem is that, under some mild assumptions on the formulation of the energy (namely being affine separable, with a sub-additive term), it is

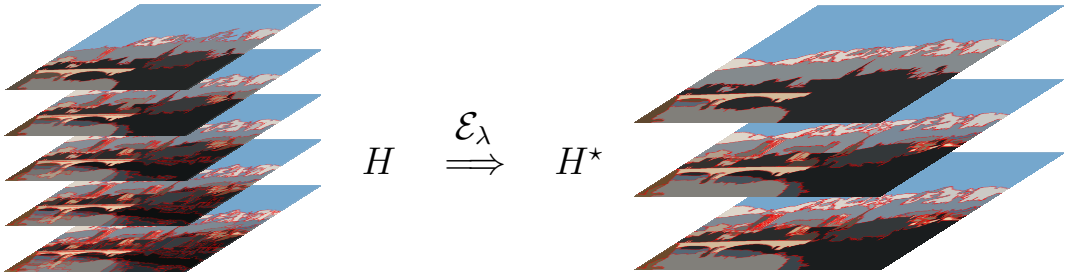


Figure 2.8: Transformation of some hierarchy H into its persistent hierarchy $H^\star = \{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ with respect to some energy $\mathcal{E}_\lambda$ parametrized by λ.

possible to extract for a hierarchy H a sequence of optimal cuts $\{\pi^\star_\lambda\}$ which are ordered by refinement. More specifically, the larger the λ, the coarser the optimal cut. Contrarily, the smaller the λ, the finer the optimal cut. The value of λ is now associated with the notion of scale of exploration of the image. As a matter of fact, it is termed the *scale parameter* in [87].

Several classical energy functions of the literature can be written as affine separable energies. It is notably the case for the piece-wise constant Mumford Shah energy as well as typical energies appearing in the MRF formulation. Often, the two competing terms are called the goodness-of-fit (GOF) (for $\mathcal{E}_\phi$) and the regularization (for $\mathcal{E}_\rho$). The former favors partitions fitting the data, thus encouraging over-segmentation in general, while the latter promotes simplicity, hence under-partition. In that context, λ acts as a trade-off between simplicity and fidelity. Using such energies, one can now analyze an image at different levels of simplicity by appropriately tuning the value of λ and conducting the energy minimization. In the case of the piece-wise constant Mumford-Shah energy, one can also remark that performing the minimization over the hierarchy allows to shift from a non-convex problem, hard to minimize, to a well-defined framework, where the global optimum can be reached easily.

2.3.2.4 Persistent hierarchy

Another consequence of the multiscale minimal cut theorem is that is it possible to assign to each element $\mathcal{R}$ of the hierarchy H two values, denoted $\lambda^+(\mathcal{R})$ and $\lambda^-(\mathcal{R})$ and called *scale of appearance* and *scale of disappearance*, respectively. They correspond intuitively to the range of values in which the region $\mathcal{R}$ is optimal (*i.e.*, when it belongs to the optimal cut):

$$\lambda^+(\mathcal{R}) \leq \lambda < \lambda^-(\mathcal{R}) \Rightarrow \mathcal{R} \in \pi^\star_\lambda \tag{2.20}$$

with the relation $\lambda^-(\mathcal{R}) = \lambda^+(\mathsf{F}(\mathcal{R}))$. A region stops being optimal when its father becomes optimal. However, nothing imposes that $\lambda^+(\mathcal{R}) \leq \lambda^-(\mathcal{R})$, meaning that a region can stop being optimal before actually starting to be optimal. Such region, which does not belong to any optimal cut of $\{\pi^\star_\lambda\}$ is said to be *non-persistent*. Conversely, a region $\mathcal{R}$ is *persistent* if $\lambda^+(\mathcal{R}) \leq \lambda^-(\mathcal{R})$. The interval $[\lambda^+(\mathcal{R}); \lambda^-(\mathcal{R})]$ is called the *interval of persistence* of $\mathcal{R}$.

(a) Original image

(b) Initial segmentation π_0

(c) $\pi^\star_\lambda$ with $\lambda \in [0.459; 0.471]$ yields 41 regions

(d) $\pi^\star_\lambda$ with $\lambda \in [0.263; 0.275]$ yields 200 regions

(e) $\pi^\star_\lambda$ with $\lambda \in [0.153; 0.154]$ yields 500 regions

Figure 2.9: Illustration of the hierarchical energy minimization framework.

The hierarchy $H^\star$, made of all persistent regions of H, is called the *persistent hierarchy*, and is composed of all the optimal cuts $\pi^\star_\lambda$ of H when λ spans $\mathbb{R}^+$. An example of such persistent hierarchy is displayed by figure 2.8. To obtain $H^\star$ in practice, the energy $\mathcal{E}_\lambda$ is seen as a function of λ, and the dynamic program is conducted over the space of such functions. The output of the dynamic program is no longer some optimal cut for a given value of λ, but some partition of $\mathbb{R}^+$ into intervals $[0, \lambda_1[\cup [\lambda_1, \lambda_2[\cup \cdots \cup [\lambda_p, +\infty[$ where all λ values within a given interval $[\lambda_i, \lambda_{i+1}[$ are leading to the same optimal cut $\pi^\star_{\lambda_i}$. The reader is referred to [87] for more practical implementation details.

Figure 2.9 illustrates this hierarchical energy minimization framework. A BPT is built over image displayed by figure 2.9a with standard parameters, namely the mean color and the Euclidean distance as region model and merging criterion, and an initial partition π_0 obtained by mean shift clustering and composed of 2156 regions (figure 2.9b). A piece-wise constant Mumford-Shah energy defined by equation (2.10) is minimized over the resulting hierarchy. The piece-wise constant Mumford-Shah energy being affine separable with sub-additive regularization term, the minimization yields a persistent hierarchy $H^\star$ composed of all the optimal cuts of H when λ spans $\mathbb{R}^+$. Some of those cuts are displayed by figures 2.9c, 2.9d and 2.9e for various values of λ. One can see that, indeed, the smaller the λ, the finer the optimal partition. In addition, all obtained partitions can be ordered by refinement (for instance, figure 2.9c is refined by figure 2.9d, in turn refined by 2.9e).

2.4 Spectral-Spatial BPT processing by means of hyperspectral unmixing

In this section, we introduce the adaptation of the BPT algorithm for hyperspectral unmixing purposes by defining a region model and merging criterion based on the induced endmembers and/or fractional abundances, and four pruning strategies based on the optimization of the spectral reconstruction error regularized by the segmentation complexity. Finally, we detail the novel methodology depicted by figure 2.11 to find an optimal segmentation of hyperspectral images from their BPT representation based on the information provided by the spectral unmixing.

2.4.1 Spectral-spatial construction of the BPT

We propose two novel region models and corresponding merging criteria based on spectral unmixing information extracted from the regions. The first one is defined by means of the spectral information provided by the endmembers induced from the regions. Thus we refer to this model as the spectral region model and merging criterion. In the second one, we propose to make use of the spatial information provided by the fractional abundances in addition to the corresponding endmembers. Therefore, we refer to this model as the spectral-spatial region model and merging criterion.

2.4.1.1 Spectral region model and merging criterion

For each region $\mathcal{R}_i$ a set of m_i endmembers $\mathbf{E}_{\mathcal{R}_i} = [\mathbf{e}_1, \ldots, \mathbf{e}_{m_i}]$ is induced by an EIA, defining the spectral region model:

$$\mathcal{M}_{\mathcal{R}_i} \stackrel{d}{=} \mathbf{E}_{\mathcal{R}_i} = [\mathbf{e}_1, \ldots, \mathbf{e}_{m_i}]\,. \tag{2.21}$$

In particular, this spectral region model is illustrated in figure 2.10 for the region labeled as $\mathcal{R}_6$ when considering only the set of endmembers $\mathbf{E}_{\mathcal{R}_6}$ locally induced over this region.

Given two neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ and $\mathbf{E}_{\mathcal{R}_i} = [\mathbf{e}_1, \ldots, \mathbf{e}_{m_i}]$, $\mathbf{E}_{\mathcal{R}_j} = \left[\mathbf{e}_1, \ldots, \mathbf{e}_{m_j}\right]$ being their respective region models, let

$$\Delta_{i,j} = [d_{kl}] = \mathcal{O}_{SAM}(\mathbf{e}_k, \mathbf{e}_l), \quad \begin{matrix} k & = 1, \ldots, m_i \\ l & = 1, \ldots, m_j \end{matrix} \tag{2.22}$$

being the $m_i \times m_j$ endmember distance matrix whose each entry d_{kl} is the spectral angle defined by equation (1.15) between endmember $\mathbf{e}_k \in \mathbf{E}_{\mathcal{R}_i}$ and $\mathbf{e}_l \in \mathbf{E}_{\mathcal{R}_j}$. The spectral merging criterion between the two regions $\mathcal{R}_i$ and $\mathcal{R}_j$ modeled by equation (2.21) is given by the spectral dissimilarity between the set of endmembers of the two regions following [84]:

$$\mathcal{O}\left(\mathcal{M}_i, \mathcal{M}_j\right) \stackrel{d}{=} d\left(\mathbf{E}_{\mathcal{R}_i}, \mathbf{E}_{\mathcal{R}_j}\right) = \|\mathbf{m}_r\|_2 + \|\mathbf{m}_c\|_2\,, \tag{2.23}$$

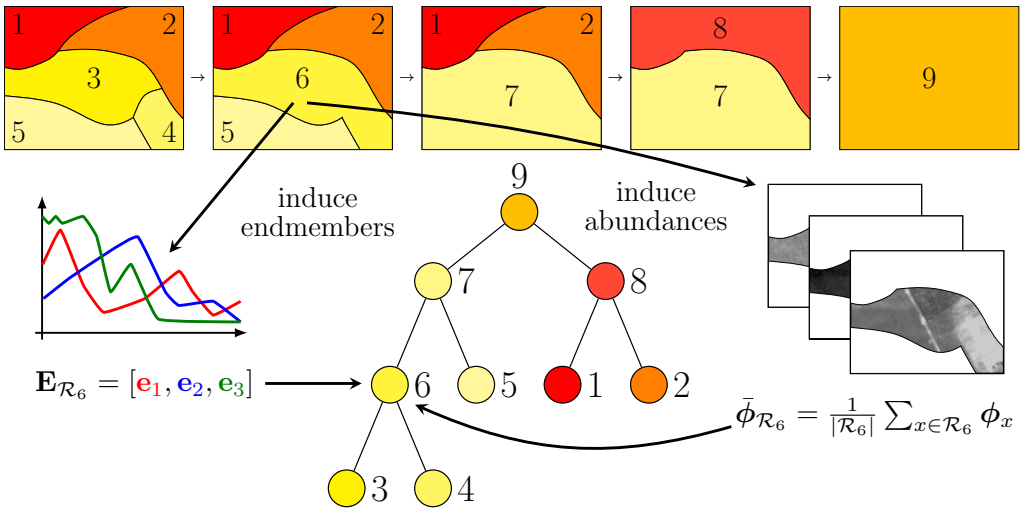


Figure 2.10: Illustration of the proposed spectral $\mathcal{M}_{\mathcal{R}_i} = \mathbf{E}_{\mathcal{R}_i}$ and spectral-spatial $\mathcal{M}_{\mathcal{R}_i} = \left(\mathbf{E}_{\mathcal{R}_i}, \bar{\phi}_{\mathcal{R}_i}\right)$ region models for the construction of the BPT.

where $\|\mathbf{m}_r\|_2$ and $\|\mathbf{m}_c\|_2$ are the minimum Euclidean norms among all row and column vectors, respectively, of the endmembers distance matrix $\Delta_{i,j}$. Once two regions merge into a new one, the set of endmembers for the new (larger) region is induced again by the given EIA. The rationale and originality of this spectral region model and merging criterion is to favor the grouping of neighboring regions that are made of similar materials (endmembers). The proposed spectral merging criterion, as it is defined, strongly penalizes regions that do not contain the same materials, therefore it is fully adapted to the underlying motivation.

2.4.1.2 Spectral-spatial region model and merging criterion

For each region $\mathcal{R}_i$ a set of m_i endmembers $\mathbf{E}_{\mathcal{R}_i} = [\mathbf{e}_1, \ldots, \mathbf{e}_{m_i}]$ is induced by some EIA, and their corresponding abundances, $\mathbf{\Phi}_{\mathcal{R}_i} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{m_i}]$, are estimated. Then, the spectral-spatial region model is defined as:

$$\mathcal{M}_{\mathcal{R}_i} \stackrel{d}{=} \left(\mathbf{E}_{\mathcal{R}_i}, \bar{\phi}_{\mathcal{R}_i}\right). \tag{2.24}$$

In the previous equation (2.24), the tuple $\left(\mathbf{E}_{\mathcal{R}_i}, \bar{\phi}_{\mathcal{R}_i}\right)$ is composed by the set of endmembers $\mathbf{E}_{\mathcal{R}_i}$ and their corresponding average fractional abundances, $\bar{\phi}_{\mathcal{R}_i} = \left[\bar{\phi}_1, \ldots, \bar{\phi}_{m_i}\right]$, with

$$\bar{\phi}_i = \frac{1}{|\mathcal{R}_i|} \sum_{x \in \mathcal{R}_i} \phi_{i,x}, \tag{2.25}$$

where $|\mathcal{R}_i|$ denotes the number of pixels in the region $\mathcal{R}_i$, and $\phi_{i,x}$ is the fractional abundance of the ith endmember for pixel $x \in \mathcal{R}_i$. An illustration of this spectral-spatial region model

Algorithm 1 Significance credits assignment algorithm.

1. $\mathcal{L} \leftarrow \{\}$
2. $\mathcal{M} \leftarrow \{(k,l):\ k=1,\ldots,m_i;\ l=1,\ldots,m_j\}$
3. Choose the minimum d_{kl} for $(k,l) \in \mathcal{M}-\mathcal{L}$
Label the corresponding (k,l) as (k',l')
4. $w_{k'l'} \leftarrow \min\left\{\bar{\phi}^i_{k'}, \bar{\phi}^j_{l'}\right\}$

if $\bar{\phi}^i_{k'} < \bar{\phi}^j_{l'}$ **then**

5. $w_{k'l} \leftarrow 0,\ \forall l \neq l'$
6. $\bar{\phi}^i_{k'} \leftarrow 0$
7. $\bar{\phi}^j_{l'} \leftarrow \bar{\phi}^j_{l'} - \bar{\phi}^i_{k'}$

else

8. $w_{kl'} \leftarrow 0,\ \forall k \neq k'$
9. $\bar{\phi}^j_{l'} \leftarrow 0$
10. $\bar{\phi}^i_{k'} \leftarrow \bar{\phi}^i_{k'} - \bar{\phi}^j_{l'}$

end if

11. $\mathcal{L} \leftarrow \mathcal{L} + \{(k',l')\}$

if $\sum_{k=1}^{m_i} \bar{\phi}^i_k > 0$ and $\sum_{l=1}^{m_j} \bar{\phi}^j_l > 0$ **then**

12. go to step 3

else

13. **return**

end if

is proposed by figure 2.10, where the region model $\mathcal{M}_{\mathcal{R}_6}$ of region $\mathcal{R}_6$ is composed of both the endmembers $\mathbf{E}_{\mathcal{R}_6}$ induced locally over $\mathcal{R}_6$ and their corresponding weighted fractional abundances $\bar{\boldsymbol{\phi}}_{\mathcal{R}_6}$.

The spectral-spatial merging criterion between two neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ modeled by equation (2.24) is given by the spectral-spatial dissimilarity between the set of endmembers and the corresponding average abundances of the two regions as it was proposed in [215]:

$$\mathcal{O}\left(\mathcal{M}_{R_i}, \mathcal{M}_{R_j}\right) \stackrel{d}{=} d\left(\left(\mathbf{E}_{\mathcal{R}_i}, \bar{\boldsymbol{\phi}}_{\mathcal{R}_i}\right), \left(\mathbf{E}_{\mathcal{R}_j}, \bar{\boldsymbol{\phi}}_{\mathcal{R}_j}\right)\right) = \sum_{k=1}^{m_i}\sum_{l=1}^{m_j} w_{kl} d_{kl}, \tag{2.26}$$

where d_{kl} is the spectral angle distance between two endmembers, $\mathbf{e}_k \in \mathbf{E}_{\mathcal{R}_i}$ and $\mathbf{e}_l \in \mathbf{E}_{\mathcal{R}_j}$, as it was defined by equation (2.22) above, and w_{kl} is a weighting coefficient measuring the significance associated to d_{kl}. The matrix of weighting coefficients, $W_{i,j} = [w_{kl}]$, $k = 1,\ldots,m_i$, $l = 1,\ldots,m_j$, is calculated using the *significance credit assignment algorithm* (see Algorithm 1) introduced in [215] which is a version of the *most similar highest priority* principle [112], where the average fractional abundances, $\bar{\boldsymbol{\phi}}_{\mathcal{R}_i}$ and $\bar{\boldsymbol{\phi}}_{\mathcal{R}_j}$ play the role of "significant credits" assigned to the spectral distances, d_{kl}. The use of the proposed spectral-spatial merging criterion promotes the merging of regions containing similar materials and in similar proportions.

2.4.2 Spectral-spatial pruning of the BPT

We present now four novel energy functions based on the spectral unmixing of the regions in the BPT representation H of an hyperspectral image. Their goal is to provide a partition, extracted from the set of all cuts $\Pi_E(H)$, which is optimal in the sense of spectral unmixing. Of course, this notion of optimality is bound to the definition of the energy term. In the following, we shall restrict to affine energies $\mathcal{E}_\lambda(\mathcal{R}) = (\mathcal{E}_\phi, \mathcal{E}_\rho) = \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})$. In a first instance, we define the composition law $\mathfrak{D}$ for the energy of a cut π as the sum over all its regions, therefore remaining in the scope of affine separable energies as they were introduced by Guigues [86, 87] and for which the theoretical results, reminded in section 2.3.2, are proved and sound. Under this framework, we propose two new energy definitions. In a second stage, we propose to use a new composition rule to express the energy of a partition, namely as the maximum of its regional energies. We first check that all theoretical results holding for separable energies export well to these new max-composed energies. Then, we propose two instances of unmixing-based max-composed energy functions.

2.4.2.1 Unmixing-based affine separable energies

The first proposed unmixing-based affine separable energy is based on the overall average RMSE, regularized by the number of regions in the partition:

$$\mathcal{E}_\lambda^{\sum\text{avg}}(\pi) = \frac{1}{|E|}\sum_{\mathcal{R}\in\pi}\sum_{x\in\mathcal{R}}\epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) + \lambda|\pi| \tag{2.27}$$

where $|E|$ and $|\pi|$ are the number of pixels in the image and the number of regions in the partition π, respectively. $\epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}})$ stands for the RMSE defined by equation (2.7), for the pixel signature $\mathbf{x}$ with respect to the estimated pixel $\hat{\mathbf{x}} = \sum_{i=1}^{m}\hat{\phi}_i\hat{\mathbf{e}}_i$, reconstructed using the set of endmembers $\hat{\mathbf{E}}_{\mathcal{R}}$ and the fractional abundances $\hat{\mathbf{\Phi}}_{\mathcal{R}}$ induced over the region $\mathcal{R}$. Defined following equation (2.27), the energy $\mathcal{E}_\lambda^{\sum\text{avg}}$ can indeed be written as an affine separable energy:

$$\mathcal{E}_\lambda^{\sum\text{avg}}(\pi) = \sum_{\mathcal{R}\in\pi}\left[\mathcal{E}_\phi^{\sum\text{avg}}(\mathcal{R}) + \lambda\mathcal{E}_\rho^{\sum\text{avg}}(\mathcal{R})\right] \text{ with } \begin{cases} \mathcal{E}_\phi^{\sum\text{avg}}(\mathcal{R}) &= \frac{1}{|E|}\sum_{x\in\mathcal{R}}\epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) \\ \mathcal{E}_\rho^{\sum\text{avg}}(\mathcal{R}) &= 1 \end{cases} \tag{2.28}$$

In (2.28), the term $\mathcal{E}_\phi^{\sum\text{avg}}$ penalizes regions whose pixels have a high reconstruction error, and can thus be seen as a goodness-of-fit (GOF) term with respect to the unmixing process. The regularization term $\mathcal{E}_\rho^{\sum\text{avg}}$ being set to 1 acts as a regularizer on the total number of regions in the partition (such regularizer was introduced in [204]). One straightforwardly check that such regularization term is sub-additive. Energy $\mathcal{E}_\lambda^{\sum\text{avg}}$ (2.27) being an affine serapable energy with sub-additive regularization term, it can therefore be minimized using the dynamic program (2.16) and (2.17) and the multiscale minimal cut theorem is guaranteed to be holding, meaning that it is possible to transform the BPT hierarchy H into its persistent version $H^\star_{\sum\text{avg}} = \{\pi^\star_\lambda\}$ where each optimal cut $\pi^\star_\lambda$ achieves a trade-off between spectral unmixing

fitting, expressed as the average RSME over the whole image, and simplicity (in terms of number of regions) controlled by the value of λ.

Similarly, we define a second unmixing-based energy, expressed as the weighted average of the maximum RMSE of the regions in the partition, regularized again by the number of regions in the partition:

$$\mathcal{E}_\lambda^{\sum \max}(\pi) = \frac{1}{|E|} \sum_{\mathcal{R}\in\pi} |\mathcal{R}| \max_{x\in\mathcal{R}} \epsilon_{\mathcal{R}}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda|\pi| \tag{2.29}$$

Energy $\mathcal{E}_\lambda^{\sum \max}$ (2.29) can be derived from energy $\mathcal{E}_\lambda^{\sum \text{avg}}$ (2.27) by replacing $\epsilon_{\mathcal{R}}(\mathbf{x}, \hat{\mathbf{x}})$ by $\max_{x\in\mathcal{R}} \epsilon_{\mathcal{R}}(\mathbf{x}, \hat{\mathbf{x}})$ in the GOF term. Therefore, all conclusions drawn for energy $\mathcal{E}_\lambda^{\sum \text{avg}}$ also hold for $\mathcal{E}_\lambda^{\sum \max}$, namely the capacity to minimize it by dynamic programming, the validity of the multiscale minimal cut theorem and the transformation of the hierarchy into its persistent version.

2.4.2.2 Max-composed energies

We now depart from the scope of affine separable energies, as they were introduced in the work of Guigues [86, 87], to focus on energies which are composed by a maximum rule $\mathfrak{D} \equiv \bigvee$, that is

$$\mathcal{E}(\pi) = \bigvee_{\mathcal{R}\in\pi} \mathcal{E}(\mathcal{R}) \ . \tag{2.30}$$

With such defined energy, the first question arising concerns the validity of the dynamic program procedure: is it possible to adapt it to handle max-composed energies? As a matter of fact, the answer is yes:

Proposition 2.1 (Minimization of a max-composed energy)
Let H be some hierarchy of partitions built over the space E. Let $\mathcal{E}$ be a max-composed energy, that is, for any $\pi \in \Pi_E$, $\mathcal{E}(\pi) = \bigvee_{\mathcal{R}\in\pi} \mathcal{E}(\mathcal{R})$. Then, for every region $\mathcal{R} \in H$, the following Bellman's dynamic program is holding:

$$\mathcal{E}^\star(\mathcal{R}) = \min \left\{ \mathcal{E}(\mathcal{R}), \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r) \right\} \tag{2.31}$$

$$\pi^\star(\mathcal{R}) = \begin{cases} \{\mathcal{R}\} & \text{if } \mathcal{E}(\mathcal{R}) \leq \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r) \\ \bigsqcup_{r\in\mathsf{S}(\mathcal{R})} \pi^\star(r) & \text{otherwise} \end{cases} \tag{2.32}$$

Proof. The proof is adapted from the one provided by Guigues [86, pp. 141-142] for separable energies. Let $\mathcal{R} \in H$ and let $H(\mathcal{R})$ be the sub-hierarchy of H rooted at $\mathcal{R}$. Define

$$\pi^\star(\mathcal{R}) = \bigsqcup_{r\in\mathsf{C}(\mathcal{R})} \pi^\star(r)$$

as the disjoint union of the optimal cuts of all the children $r \in \mathsf{C}(\mathcal{R})$ of $\mathcal{R}$. Similarly, let $\pi'(\mathcal{R})$ be another partition partition of $\mathcal{R}$. π' can be written as the disjoint union of some cuts $\pi'(r)$ of the children r of $\mathcal{R}$,

$$\pi'(\mathcal{R}) = \bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi'(r).$$

$\mathcal{E}$ being max-composed, one has $\mathcal{E}(\pi^\star(\mathcal{R})) = \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}(\pi^\star(r))$ and $\mathcal{E}(\pi'(\mathcal{R})) = \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}(\pi'(r))$ and since $\pi^\star(r)$ is the optimal cut in $H(r)$, we have

$$\begin{aligned} \mathcal{E}(\pi^\star(r)) = \mathcal{E}^\star(r) \leq \mathcal{E}(\pi'(r)) \quad \forall r \in \mathsf{C}(\mathcal{R}) \\ \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r) \leq \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}(\pi'(r)) \\ \mathcal{E}^\star(\mathcal{R}) \leq \mathcal{E}(\pi'(\mathcal{R})) \end{aligned}$$

In conclusion,

if	$\mathcal{E}(\mathcal{R}) \leq \mathcal{E}(\pi^\star(\mathcal{R}))$, then	$\pi^\star(\mathcal{R}) = \{\mathcal{R}\}$		and	$\mathcal{E}^\star(\mathcal{R}) = \mathcal{E}(\mathcal{R})$
otherwise		$\pi^\star(\mathcal{R}) = \bigsqcup_{r\in\mathsf{C}(\mathcal{R})} \pi^\star(r)$		and	$\mathcal{E}^\star(\mathcal{R}) = \bigvee_{r\in\mathsf{C}(\mathcal{R})} \mathcal{E}^\star(r)$

□

Still following the approach of Guigues, we now focus on affine max-composed energies, namely energies which can be written as

$$\mathcal{E}_\lambda(\pi) = \bigvee_{\mathcal{R}\in\pi} \left[\mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})\right] . \tag{2.33}$$

The next question arising concerns the validity of the multiscale minimal cut theorem. When working with affine separable energies, Guigues used in his proof [86, pp. 161-162] the linearity of the sum operator (*i.e.*, the fact that $\sum_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R}) = \sum_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda \sum_{\mathcal{R}\in\pi} \mathcal{E}_\rho(\mathcal{R})$) combined with the sub-additivity condition on $\mathcal{E}_\rho$ (namely, $\mathcal{E}_\rho(\mathcal{R}) \leq \sum_{\mathcal{R}'\in\pi(\mathcal{R})} \mathcal{E}_\rho(\mathcal{R}')$ for some partition partition $\pi(\mathcal{R})$ of $\mathcal{R}$). Using the maximum operator $\bigvee$ in our case, which is not linear, we cannot directly adapt the proof of Guigues as it was done for the dynamic program. In his work, Kiran [101, p. 53] introduced the notion of inf-modularity as a generalization of sub-additivity, and proved that the multiscale minimal cut theorem was holding for any family of the type $\mathcal{E}_\lambda(\pi) = \mathcal{E}_\phi(\pi) + \lambda\mathcal{E}_\rho(\pi)$ when the term $\mathcal{E}_\rho$ is inf-modular, which is still not the present case. However, as suggested in [101], it is nevertheless possible to prove the validity of the multiscale minimal cut theorem for energies with other composition rules, based on the monotonicity of the mapping $\lambda \mapsto \mathcal{E}_\lambda(\pi)$. Using this argument, we can formulate again the multiscale minimal cut theorem for max-composed energies:

Theorem 2.2 (Multiscale minimal cut for max-composed energies)
Let H be a hierarchy on a set E, and let $\mathcal{E}_\lambda(\pi) = \bigvee_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})$ be an affine max-composed energy such that $\mathcal{E}_\rho(\mathcal{R}) \geq 0$ $\forall \mathcal{R} \in H$. Then the family of optimal cuts $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ can be ordered by refinement, i.e.:

$$\forall \lambda_1, \lambda_2, 0 \leq \lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2} \tag{2.34}$$

Proof. The proof is given in appendix A. □

Therefore, max-composed affine energies can be processed in the same way as separable affine energies in terms of minimization by dynamic program and transformation of a hierarchy H into its persistent version $H^\star$.

2.4.2.3 Unmixing-based max-composed energies

Following the previous theoretical results, we now propose to define unmixing-based energies composed by the maximum law as defined by (2.33):

$$\mathcal{E}_\lambda(\pi) = \bigvee_{\mathcal{R}\in\pi} [\mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})]$$

where $\mathcal{E}_\phi$ and $\mathcal{E}_\rho$ play the role of bounds on data fitting and complexity, respectively. The optimal partition $\pi^\star_\lambda$ with respect to such energy is the one that minimize the regularized combination of both bounds.

The first proposed max-composed unmixing-based energy function is defined as

$$\mathcal{E}_\lambda^{\bigvee\,\mathrm{avg}}(\pi) = \bigvee_{\mathcal{R}\in\pi} \left[\frac{1}{|\mathcal{R}|}\sum_{x\in\mathcal{R}} \epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) + \frac{\lambda}{|\mathcal{R}|}\right] \tag{2.35}$$

where the GOF term $\mathcal{E}_\phi(\mathcal{R})$ is expressed in terms of average RMSE within the region $\mathcal{R}$, and the regularization $\mathcal{E}_\rho(\mathcal{R})$ is defined as the inverse of the region size. The optimal cut of this energy minimizes the upper bound on the average reconstruction error of the regions and at the same time maximizes the lower bound on the size of the regions in the partition, with λ acting as a trade-off parameter and thus giving more weight to one bound or the other.

Finally, we define a last unmixing-based energy function by replacing the average RMSE of region $\mathcal{R}$ in $\mathcal{E}_\lambda^{\bigvee\,\mathrm{avg}}$ (2.35) by the maximum RMSE:

$$\mathcal{E}_\lambda^{\bigvee\,\max}(\pi) = \bigvee_{\mathcal{R}\in\pi} \left[\max_{x\in\mathcal{R}} \epsilon_{\mathcal{R}}(\mathbf{x},\hat{\mathbf{x}}) + \frac{\lambda}{|\mathcal{R}|}\right] \tag{2.36}$$

Energy $\mathcal{E}_\lambda^{\bigvee\,\max}$ has for minimizer a partition that minimizes the upper bound on the maximal reconstruction errors and at the same time maximizes the lower bound of the size of the regions in the partition, and can be seen as more restrictive version than $\mathcal{E}_\lambda^{\bigvee\,\mathrm{avg}}$ as regions having an overall low RMSE with a single badly reconstructed pixel will be more penalized in $\mathcal{E}_\lambda^{\bigvee\,\max}$.

2.4.2.4 Use of a size constraint

It is sometimes interesting to constrain the set of valid partitions, $\Pi_E(H)$, to those whose all regions size is above a given minimum size. For instance, the segmentation of the image could

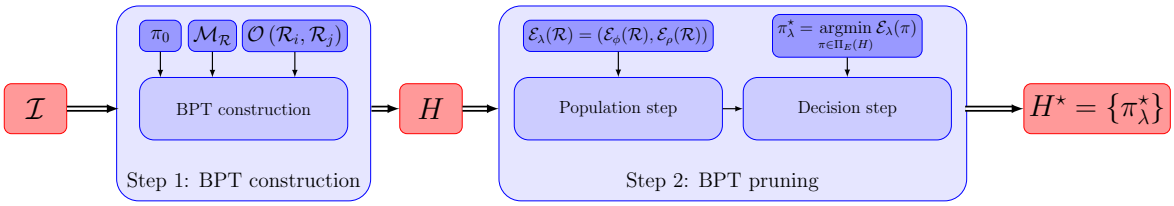


Figure 2.11: Flowchart of the proposed novel methodology.

be later used for applications that require a minimum number of pixels to work (to estimate some statistical parameters for instance). In these cases, the set $\Pi_E(H)$ of valid partitions in the formulation of the optimization problems is replaced by the subset of size-constrained valid partitions, $\Pi_E^c(H)$:

$$\Pi_E^c(H) = \{\pi \in \Pi_E(H), \text{ s.t. } \forall \mathcal{R} \in \pi, |\mathcal{R}| \geq c\}\,, \tag{2.37}$$

where $|\mathcal{R}|$ denotes the number of pixels in region $\mathcal{R}$ and $c \geq 0$ is a threshold on the region size. If $c = 0$, the term (2.37) has no effect and the pruning criterion is considered to be unconstrained.

2.4.3 Proposed methodology

Fig. 2.11 shows the flow diagram of the proposed general methodology to obtain an optimal segmentation from a hyperspectral image, by pruning the BPT representation of the image using the information provided by the spectral unmixing process. The procedure, decomposed in two steps, is as follows:

Step one. First, a BPT representation H of the input hyperspectral image $\mathcal{I}$ is obtained. In order to build the BPT, one must provide three input parameters, namely the initial partition of the image π_0, a region model $\mathcal{M}_\mathcal{R}$ and an associated merging criterion $\mathcal{O}(\mathcal{R}_i, \mathcal{R}_j)$. For the initial partition, the only constraint is that it provides an under-segmentation of the image, with initial regions small enough not to encompass "actual" regions, and accurate enough to be able to reconstruct those regions with a good accuracy. For the region model and associated merging criterion, we propose to use either, the spectral region model (2.21) and merging criterion (2.23) or the spectral-spatial region model (2.24) and merging criterion (2.26) previously defined. In order to do that, a spectral unmixing process is run independently for each region $\mathcal{R}$ (see figure 2.12). First, the virtual dimensionality $\delta_\mathcal{R}$ of the region $\mathcal{R}$ is computed using the Hyperspectral Signal Subspace Estimation (Hysime) algorithm [22]. The value of $\delta_\mathcal{R}$ works as an estimation of the number m of endmembers present in the region. If the region is too small to correctly estimate the number of endmembers (due to the presence of close to singular covariance matrices during the application of the Hysime algorithm), that is, if $\delta_\mathcal{R} = 0$ or $\delta_\mathcal{R} > |\mathcal{R}|$, being $|\mathcal{R}|$ the number of pixels in the region, then its region model $\mathcal{M}_\mathcal{R}$ is set to the mean spectrum of the region $\mathcal{M}_\mathcal{R} = \boldsymbol{\mu}_\mathcal{R}$. This happens in very small and

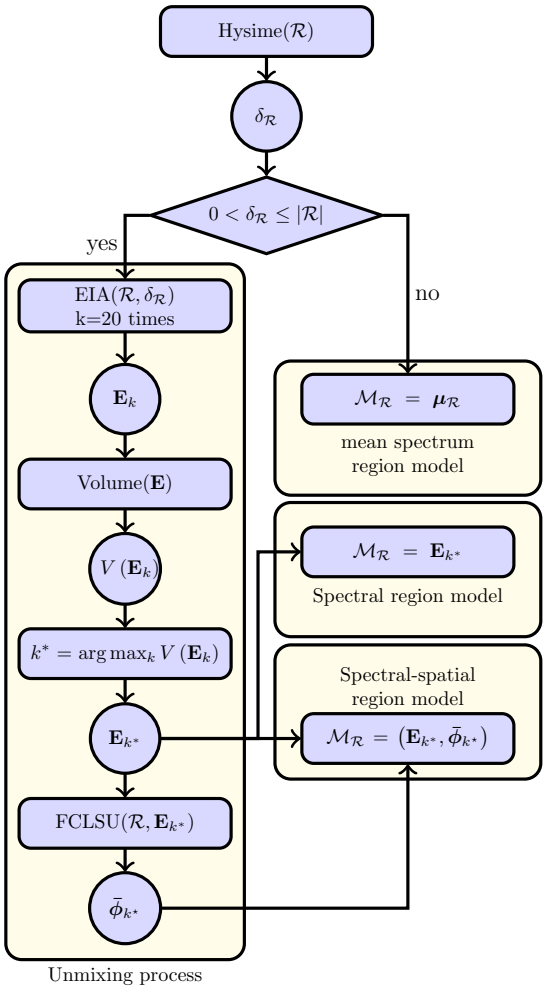


Figure 2.12: Workflow of the spectral unmixing process to obtain the region model.

homogeneous regions, so the mean spectrum $\boldsymbol{\mu}_{\mathcal{R}}$ acts as a single endmember. Otherwise, an EIA is run over the $|\mathcal{R}|$ pixels of the region to induce the corresponding set of endmembers. To overcome the stochastic part of most of the EIAs, the induction algorithm is run a number of times k for each region, and the set of endmembers yielding the larger simplex volume [227], $V_k(\mathbf{E})$, among the k trials is retained. If the spectral region model is selected to build the BPT representation, the region model is defined by these endmembers as it is described in (2.21). If the spectral-spatial region model is selected, the FCLSU is conducted and the fractional abundances of the induced endmembers are estimated for each pixel in the region. The region model is then defined by the endmembers and their average fractional abundances as it is described in (2.24). Computing the unmixing information for each region $\mathcal{R}$ during the construction of the BPT can be computationally expensive, but once the BPT has been populated for this information, it can be stored and any posterior processing of the BPT representation becomes very fast. This trade-off is common in the analysis of images by means of tree-based representations.

Step two. The second step features the pruning of the BPT H to transform it into its persistent version $H^\star$. This, as any pruning operation, is done in two steps:

1. The population step involves the computation of the energy of each region $\mathcal{R}$. Here, we propose to work with affine energies, namely energies being the sum of a GOF term $\mathcal{E}_\phi(\mathcal{R})$ and a regularization term $\mathcal{E}_\rho(\mathcal{R})$ weighted by a coefficient λ. Following the novel energy definitions introduced in subsection 2.4.2, the regularization term requires, at most, the region size $|\mathcal{R}|$. On the other hand, the GOF term is based on the reconstruction error of each region, calculated by the RMSE (2.7), given the set of endmembers and corresponding abundances. Note that, if any of the two proposed unmixing-based region models are used to build the BPT representation, this information is already stored during the construction and can be used as is. If not, for instance, when using a mean spectrum region model, the spectral unmixing process defined above should be run for each region in order to induce the endmembers and estimate the fractional abundances.
2. Then, given a composing law for the energy being either $\mathfrak{D} = \sum$ or $\mathfrak{D} = \bigvee$, one can then define the energy $\mathcal{E}(\pi)$ of a cut $\pi \in \Pi_E(H)$ and seek for the optimal one given a value of λ. Even better, if the definition of the enegy allows it, one can directly compute all optimal cuts by viewing the energy as a function of λ and conducting the dynamic program (2.17) and (2.16) over the space of such function (as advocated in [87]), producing the persistent hierarchy $H^\star$.

2.5 Experimental methodology

2.5.1 Hyperspectral datasets

We propose to use in the experiments two real hyperspectral data sets. Their selection is supported by the fact that these scenes have been widely used to validate hyperspectral segmentation and spectral unmixing applications, and currently constitute benchmarks used to validate new algorithms thanks to the availability of reliable reference information. The considered scenes can be summarized as follows:

The Pavia University hyperspectral image. It was collected by the ROSIS-03 sensor over the facilities of the University of Pavia in Italy. After discarding pixels with no information and noisy spectral bands, the image has a spatial size of 610×340 pixels with a spatial resolution of 1.3 m per pixel, and 93 spectral bands comprised in the range of 430-860 nm. Figure 2.13a features a false color representation of the Pavia University scene. The scene shows an urban area comprised of different buildings, parking lots, roads and other typical human-made constructions, together with trees, green areas and bare soil.

The Cuprite hyperspectral scene. It was acquired by the NASA's AVIRIS sensor [85] and covers the Cuprite mining district in western Nevada, USA. This sensor collects data in 224 contiguous spectral bands with a bandwidth of 0.10 μm in the range of $0.4 - 2.5 \mu m$. 200

(a) (b)

Figure 2.13: False color representation of (a) the Pavia University scene and (b) the Cuprite scene.

bands remain after removing noisy bands due to atmospheric water absorption. Each pixel represents a $20m^2$ square cell. The data used in the experiments is a 250×190 subset of the original scene covering the mineralogical region of interest. Figure 2.13b shows a false color representation of the subset of the Cuprite scene used to conduct the experiments. The scene is well-known and widely used in hyperspectral community thanks to the extensive reference information available for this scene from the United States Geological Survey (USGS) [6].

2.5.2 Experimental methodology

This section describes the procedure adopted to conduct the analysis of the two aforementioned hyperspectral scenes. Specifically, we describe the steps followed in order to build the BPT representations and extract some optimal cuts, as well as the quantitative measures employed to assess the quality of these optimal cuts.

For each dataset, we build three independent BPT representations, each using a specific region model and merging criterion:

6. http://speclab.cr.usgs.gov/cuprite.html

- one BPT using the mean spectrum region model (1.13) and associated spectral angle (1.15), hereafter denoted H_μ.
- one BPT using the proposed spectral distance model (2.21), based on the endmembers induced over each region, and the associated proposed spectral distance (2.23), H_e standing for this configuration.
- one BPT using the proposed spectral-spatial distance model (2.24), based on the endmembers induced over each region and their corresponding average abundances, and the associated proposed spectral-spatial distance (2.26). This BPT will be denoted $H_{e\bar{\phi}}$.

In all cases, the BPT is built over an initial partition π_0 of the image, obtained by the hyperspectal watershed [188] method using a multidimensional morphological gradient [146]. This method produces severely over-segmented partitions maps and has already shown to be relevant for the hierarchical representation of hyperspectral images [200, 216]. In addition, the priority term [36] enforcing small regions to merge with priority during the merging process is set to 15%. Each BPT representation is then populated with the endmembers and fractional abundances from an unmixing process run in each node, as explained in section 2.4.3 (note that this information is already available for H_e and $H_{e\bar{\phi}}$). The Vertex Component Analysis (VCA) algorithm [144] is chosen to induce the endmembers. Due to the stochasticity of such algorithm, several runs are made for each region of the BPT (20 independent runs in the present case), and the set of endmembers yielding the simplex with maximal volume is retained.

Then, each BPT representation H_μ, H_e and $H_{e\bar{\phi}}$ is pruned by minimizing the four proposed unmixing-based energy function $\mathcal{E}_\lambda^{\sum\text{avg}}$, $\mathcal{E}_\lambda^{\sum\max}$, $\mathcal{E}_\lambda^{\bigvee\text{avg}}$ and $\mathcal{E}_\lambda^{\bigvee\max}$ to generate, in each case, a set of optimal cuts whose number of regions matches some predefined numbers. In particular:

- For the Pavia University scene, we extract 8 optimal cuts having $5, 10, 40, 75, 100, 225, 350$ and 850 regions (or the cuts having a number of regions as close as possible from the desired numbers).
- The same procedure is conducted for the Cuprite scene, with expected numbers of regions being set to $5, 10, 20, 35, 50, 75, 150$ and 500.

In both cases, the desired numbers of regions were arbitrarily chosen. Finding the correct value of λ yielding the optimal cut with the appropriate number of regions may seem to be a tedious task. In practice however, there is a nice workaround to this issue: each energy function allows to easily compute at once all the optimal cuts of $H_*, * = \{\mu, e, e\bar{\phi}\}$ when λ spans $\mathbb{R}^+$, producing the persistent hierarchy $H_*^\star$ by stacking all cuts. Each optimal cut $\pi_\lambda^\star \in \Pi_E(H_*)$ corresponds to a horizontal cut of $H_*^\star$. Therefore, instead of looking for the correct value of λ producing the optimal cut of H_* with an appropriate number of regions, one can simply browse the horizontal cuts of $H_*^\star$ and stop when one with the desired number of regions is found.

In addition to the four proposed energy functions, each BPT is also pruned by two additional strategies:

- The horizontal cut producing the partition with the desired (or as close as possible) number of regions.
- The optimal cut with respect to the energy function proposed by Valero in [204] and

defined as an affine separable energy

$$\mathcal{E}_\lambda^{\sum \text{SID}}(\pi) = \sum_{\pi \in \mathcal{R}} \mathcal{E}_\phi(\mathcal{R}) + \lambda|\pi| \tag{2.38}$$

where the GOF term $\mathcal{E}_\phi$ is defined as

$$\mathcal{E}_\phi(\mathcal{R}) = \sum_{x \in \mathcal{R}} \mathcal{O}_{SID}(\mathbf{x}, \boldsymbol{\mu}_\mathcal{R}) + \begin{cases} 0 \text{ if } |\mathcal{R}_l| \text{ and } |\mathcal{R}_r| \leq \tau \\ \sum_{x \in \mathcal{R}_l} \mathcal{O}_{SID}(\mathbf{x}, \boldsymbol{\mu}_{\mathcal{R}_r}) + \sum_{x \in \mathcal{R}_r} \mathcal{O}_{SID}(\mathbf{x}, \boldsymbol{\mu}_{\mathcal{R}_l}) \text{ otherwise} \end{cases} \tag{2.39}$$

where $\mathcal{O}_{SID}$ denotes the spectral information divergence measure, as defined by (1.17), $\boldsymbol{\mu}_\mathcal{R}, \boldsymbol{\mu}_{\mathcal{R}_l}$ and $\boldsymbol{\mu}_{\mathcal{R}_r}$ are the mean spectra of regions $\mathcal{R}$, $\mathcal{R}_l$ and $\mathcal{R}_r$, the latter two being the left and right children of $\mathcal{R}$. The first term of (2.39) measures the error committed when replacing all pixel spectra in region $\mathcal{R}$ by their mean value $\boldsymbol{\mu}_\mathcal{R}$, thus penalizing spectrally inhomogeneous regions. The second term evaluates the error of replacing each pixel spectrum of the child region $\mathcal{R}_l$ by the mean spectrum of its sibling $\mathcal{R}_r$ and vice versa, in order to regularize the case where the region $\mathcal{R}$ has a child which is much larger than the other one (the contribution of the small child being negligible in the first error term, even if spectrally different from $\boldsymbol{\mu}_\mathcal{R}$). In practice, the second term is added to $\mathcal{E}_\phi(\mathcal{R})$ if the two children have a size greater than a predefined threshold τ (set to 3 pixels in [204]) in order to make this estimation reliable.
The energy (2.38) being affine separable with a sub-additive regularization term (as $\mathcal{E}_\rho(\mathcal{R}) = 1$), the optimal cuts with the desired number of regions are extracted from the three BPTs H_μ, H_e and $H_{e\bar{\phi}}$ in the exact same fashion as for the four proposed unmixing-based energies.

In order to quantitatively compare the segmentations obtained by the different pruning strategies, we compare the original hyperspectral image $\mathbf{X}$ to the one obtained by the unmixing reconstruction, $\hat{\mathbf{X}} = \hat{\mathbf{E}}\hat{\mathbf{\Phi}}$, calculated from the partitions obtained by the different BPT representation models, pruning criteria and expected partition sizes. The reconstruction $\hat{\mathbf{X}}$ is made piece-wise, where the endmembers and fractional abundances obtained in each region $\mathcal{R}$ of a given segmentation π are used to reconstruct only the pixels within this region. We propose to use of four different image reconstruction quality measures:

– The average RMSE measure the average Euclidean error between $\mathbf{X}$ and $\hat{\mathbf{X}}$

$$\text{avgRMSE}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \frac{1}{|E|} \sum_{x \in E} \epsilon\left(\mathbf{x}, \hat{\mathbf{x}}\right) \tag{2.40}$$

by averaging the RMSE $\epsilon(\mathbf{x}, \hat{\mathbf{x}})$ (2.7) of each reconstructed spectrum $\hat{\mathbf{x}}$ with respect to the true one $\mathbf{x}$ over the number of pixels $|E|$ in the image. If $\mathbf{X}$ is perfectly reconstructed, then $\text{avgRMSE}(\mathbf{X}, \bar{\mathbf{X}}) = 0$

– The average spectral angle error (SAE) is similar to the average RSME, but measures instead the average angular error between $\mathbf{X}$ and $\hat{\mathbf{X}}$

$$\text{avgSAE}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \frac{1}{|E|} \sum_{x \in E} \text{SAD}\left(\mathbf{x}, \bar{\mathbf{x}}\right) \tag{2.41}$$

where $\mathrm{SAD}(\mathbf{x}, \bar{\mathbf{x}})$ is the spectral angle (1.15) distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$. Similarly, a perfectly reconstructed image yields $\mathrm{avgSAE}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = 0$.

– The average Q index measures the correlation between the the original $\mathbf{X}$ and the reconstructed $\hat{\mathbf{X}}$ images

$$\mathrm{avgQ}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \frac{\sigma_{\mathbf{X}\hat{\mathbf{X}}}}{\sigma_{\mathbf{X}}\sigma_{\hat{\mathbf{X}}}} \times \frac{2\boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\mu}_{\hat{\mathbf{X}}}}{\|\boldsymbol{\mu}_{\mathbf{X}}\|_2^2 + \|\boldsymbol{\mu}_{\hat{\mathbf{X}}}\|_2^2} \times \frac{2\sigma_{\mathbf{X}}\sigma_{\hat{\mathbf{X}}}}{\sigma_{\mathbf{X}}^2 + \sigma_{\hat{\mathbf{X}}}^2}, \tag{2.42}$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\hat{\mathbf{X}}}$ are the mean N-dimensional vectors of the original and reconstructed images respectively, $\sigma_{\mathbf{X}}$ and $\sigma_{\hat{\mathbf{X}}}$ denote the variances, and $\sigma_{\mathbf{X}\hat{\mathbf{X}}}$ the covariance. $\mathrm{avgQ}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = 1$ for an ideal image reconstruction.

– The ERGAS (Erreur Relative Globale Adimensionelle de Synthèse) quality measure, which evaluates both spectral and spatial divergences:

$$\mathrm{ERGAS}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = 100\sqrt{\frac{1}{|E|}\sum_{x \in E}\left(\frac{\epsilon(\mathbf{x}, \hat{\mathbf{x}})}{\mu_{\mathbf{x}}}\right)^2}, \tag{2.43}$$

where $\mu_{\mathbf{x}}$ denotes the mean (scalar) value of pixel spectrum $\mathbf{x}$. The lower the ERGAS value, the better reconstructed the image.

With the previously defined four quality measures, one can then assess how well reconstructed (from a spectral point of view for measures (2.40) and (2.41), and from a spectral-spatial point of view for the two others (2.42) and (2.43) measures) are the images, based on the obtained segmentations. The proposed four unmixing-based energy functions were intended to produce a segmentation of the image being optimal with respect to the unmixing reconstruction error, and should therefore lead to average RMSE, average SAD and ERGAS values as low as possible, and an average Q index close to 1.

2.6 Results

2.6.1 Pavia University data set

2.6.1.1 Reconstruction errors

Figures 2.14, 2.15 and 2.16 show the quantitative reconstruction quality measures of the different pruning strategies applied over the BPT representations H_μ (mean spectrum region model), H_e (proposed spectral region model) and $H_{e\bar{\phi}}$ (proposed spectral-spatial region model) of the Pavia University scene, respectively. Each point in the plots represents a partition obtained by each of the pruning strategies over the corresponding BPT. In order to compare them, we plot the quality measure with respect to the number of regions contained in each partition.

Several observations arise when analyzing the curves. First of all, the four proposed unmixing-based energies $\mathcal{E}_\lambda^{\sum\,\mathrm{avg}}, \mathcal{E}_\lambda^{\sum\,\max}, \mathcal{E}_\lambda^{\bigvee\,\max}$ and $\mathcal{E}_\lambda^{\bigvee\,\mathrm{avg}}$ lead to optimal cuts that out-

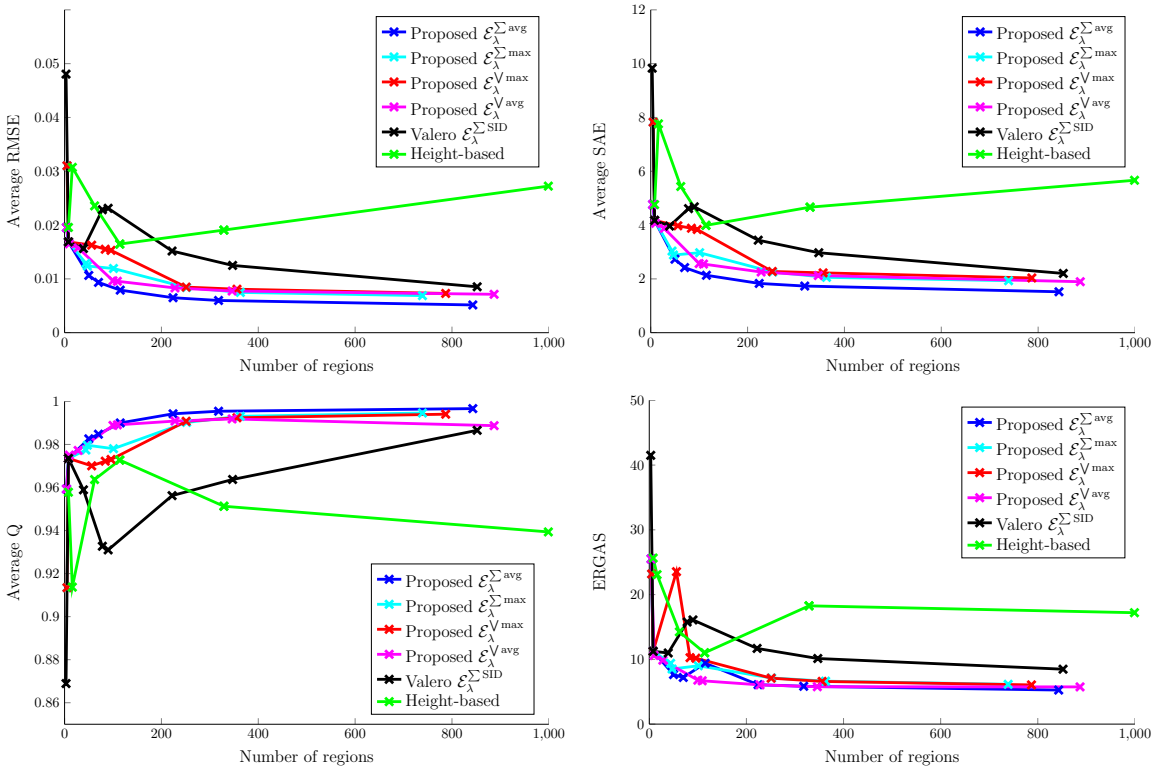


Figure 2.14: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation H_μ (mean spectrum region model) of Pavia University image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

perform height-based and Valero $\mathcal{E}_\lambda^{\sum\text{SID}}$ optimal partitions in all cases. This phenomenon can be easily interpreted: the latter two strategies are not designed to produce partitions with low reconstruction errors, as the height-based pruning only depends on the merging order of the regions during the construction of the BPT while $\mathcal{E}_\lambda^{\sum\text{SID}}$ produces partition with spectrally homogeneous regions. It is more delicate to evaluate the relative performances of the proposed unmixing-based energy functions, as the differences are rather small. However, the two affine separable energies seem to perform slightly better than their max-composed counterparts. A possible explanation is that $\mathcal{E}_\lambda^{\sum\text{avg}}$ and $\mathcal{E}_\lambda^{\sum\text{max}}$ aims at finding partitions with a low overall average RMSE for the former, and a low weighted average of maximum RMSE for the latter, with both tend to produce an overall low RSME over the whole image. The two max-composed energies $\mathcal{E}_\lambda^{\bigvee\text{avg}}$ and $\mathcal{E}_\lambda^{\bigvee\text{max}}$ operate differently, as they minimize the bound on the mean and maximum RMSE of all regions of the partition, but may lead to more averagely reconstructed pixels in the whole image, hence higher mean reconstruction errors for the whole image. The quality measures being computed over the whole image (and not region-wise) may also favor separable energies over max-composed ones. One can also remark that, for all cases but one, the quality measures have an overall decreasing behavior (expect

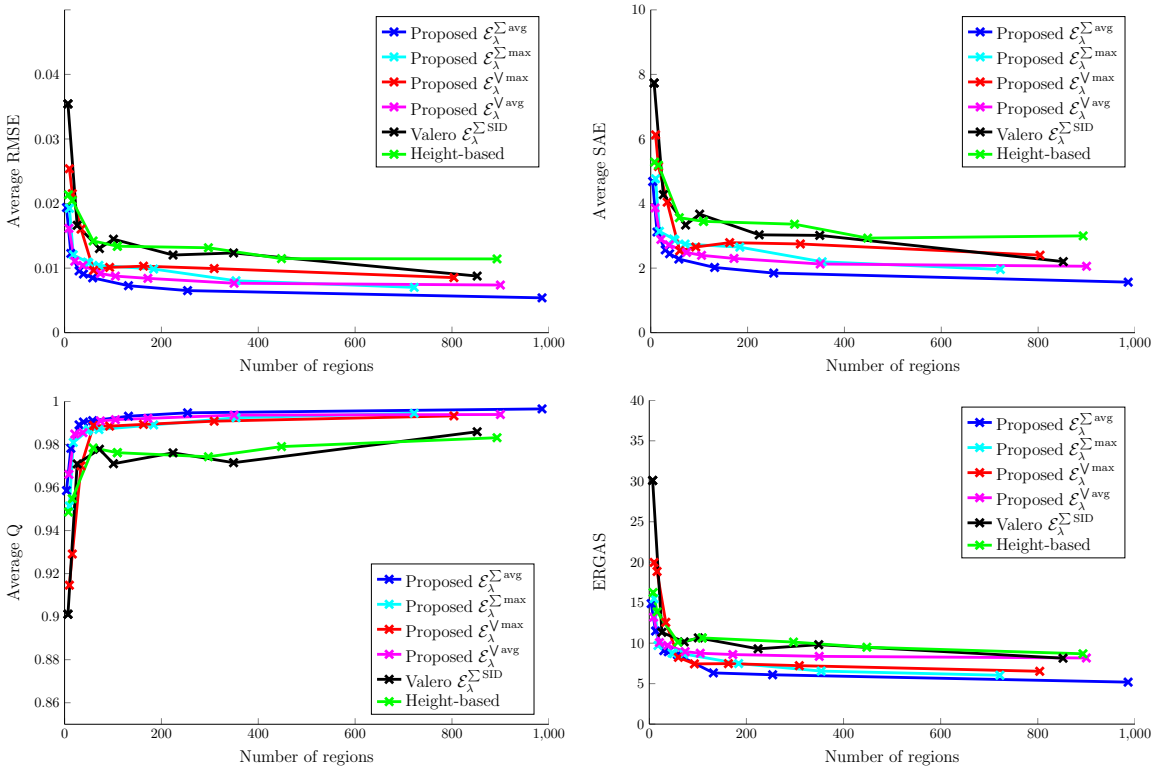


Figure 2.15: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation H_e (proposed spectral-based region model) of Pavia University image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

for the average Q which is globally increasing since the closer to 1, the better in this case) with respect to the number of regions in the partition. The only exception concerns the BPT H_μ (whose construction is based on the mean spectrum region model) when pruned with the height-based approach. In this special combination, both the construction and pruning of the BPT are unrelated to the desired goal being a segmentation optimal in terms of spectral unmixing. Going farther, one can see that the height-based approach as well as energy $\mathcal{E}_\lambda^{\sum \text{SID}}$ perform better on H_e and $H_{e\bar{\phi}}$, where the unmixing information has been taken into account during the construction of the hierarchy, than on H_μ. It confirms that building the BPT in a appropriate way with respect to the task is absolutely relevant and necessary in order to achieve the intended application, even if the pruning strategy is not fully adapted to the desired goal.

2.6.1.2 Segmentation results

Figure 2.17 shows the optimal cuts with respect to all five energies, namely (from left to right) $\mathcal{E}_\lambda^{\sum \text{avg}}, \mathcal{E}_\lambda^{\sum \max}, \mathcal{E}_\lambda^{\bigvee \max}, \mathcal{E}_\lambda^{\bigvee \text{avg}}$ and $\mathcal{E}_\lambda^{\sum \text{SID}}$, for the spectral-spatial BPT representation

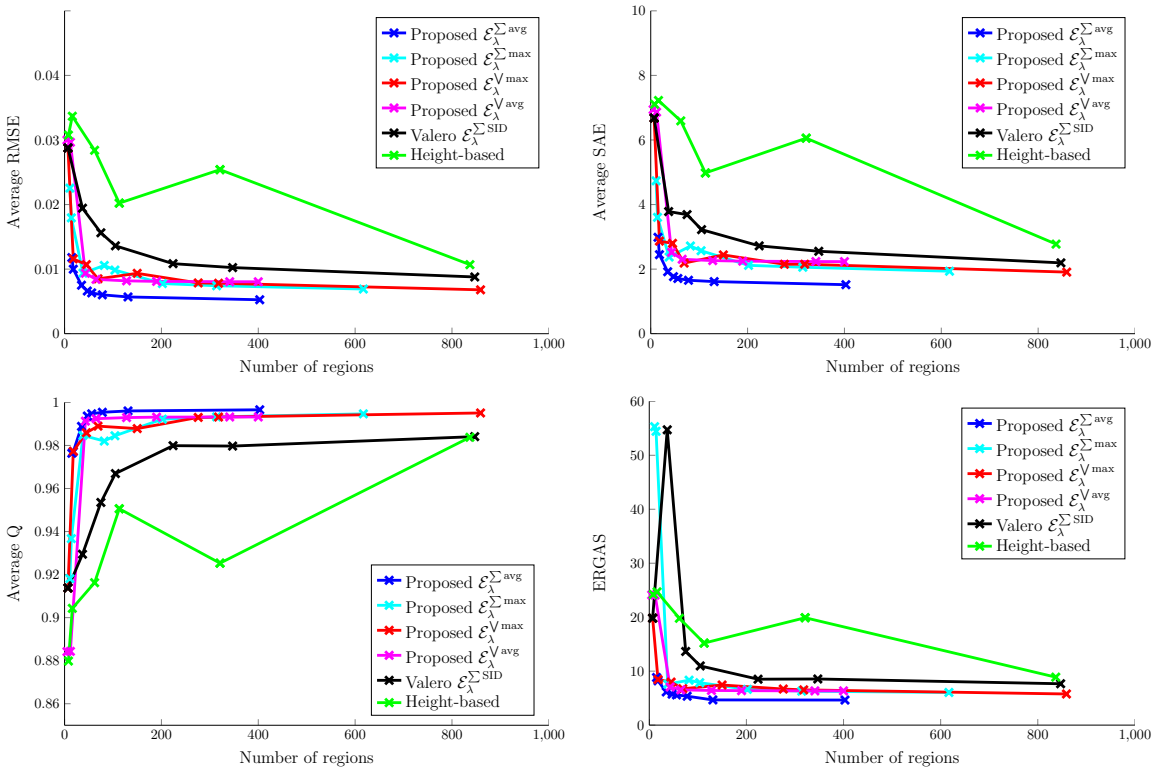
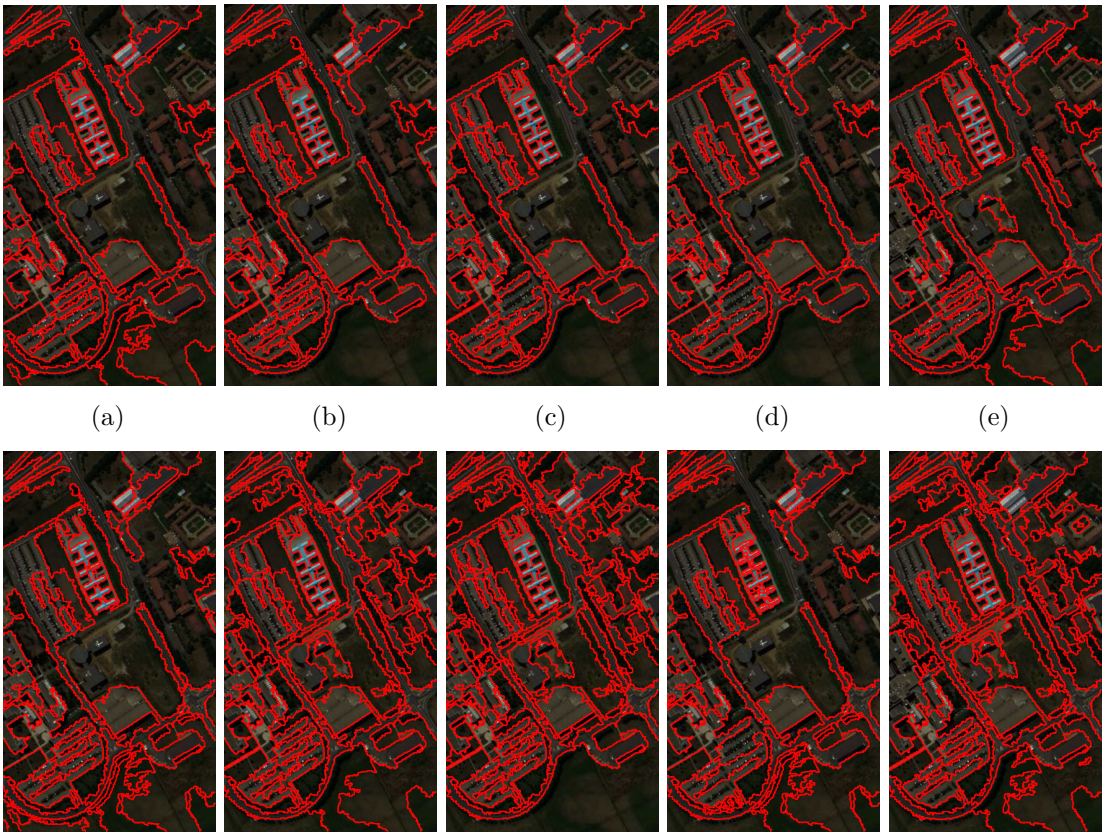


Figure 2.16: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation $H_{e\bar{\phi}}$ (proposed spectral-spatial based region model) of Pavia University image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

$H_{e\bar{\phi}}$ of the Pavia University scene. The top row shows the optimal partitions with (or close to) 50 regions, while the bottom row shows the optimal partitions with (or close to) 100 regions. The first comment that can be made is that, in both cases, the resulting partition are strongly under-segmented. As a matter of fact, when looking at figure 2.13a, one can see that the scene is composed of a multitude of regions of interest (in the sense that they bear some semantic meaning), such as the various buildings, the parking lots, the roads, the grassy areas and so on. However, in order to analyze the influence of the used spectral unmixing information on the resulting partitions, one must examine large enough regions (as it is recalled that, during the construction of a BPT, the estimated intrinsic dimensionality of each region is used to define the region model. If the region is too small, notably, then the region model is set to the mean spectrum, assumed to be the single endmember). Therefore, the evident under-segmentation is not considered to be an issue in the present case.

Among all five energy functions, $\mathcal{E}_\lambda^{\sum\text{SID}}$ is the one that seem to correctly segment the most visually salient regions. This is explained by the formulation of the energy: as pointed out in [204], it aims at producing segmentations with spectrally homogeneous regions, and is thus more adapted to the design of partitions that match the visual perception. It is a little



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 2.17: Optimal cuts extracted from the BPT representation $H_{e\bar{\phi}}$ of Pavia University scene by minimizing: (a)(f) $\mathcal{E}_\lambda^{\sum \text{avg}}$, (b)(g) $\mathcal{E}_\lambda^{\sum \text{max}}$, (c)(h) $\mathcal{E}_\lambda^{\bigvee \text{max}}$, (d)(i) $\mathcal{E}_\lambda^{\bigvee \text{avg}}$, and (e)(j) $\mathcal{E}_\lambda^{\sum \text{SID}}$. Top row segmentations have around 50 regions, bottom row segmentations have around 100 regions.

bit harder to interpret the optimal partitions of the proposed unmixing-based energies. The obtained regions are in fact supposed to be optimal with respect to the reconstruction error, which cannot be interpreted visually as it is related to the endmembers and abundances of each region. Nevertheless, one can still see some correctly delineated structures (some grass, roads, parking lots or building), especially for the segmentation displayed by the second row of figure 2.17. In addition, there are not many visual differences among the four proposed approaches, the only noticeable one being that the optimal cuts of $\mathcal{E}_\lambda^{\sum \text{avg}}$ and $\mathcal{E}_\lambda^{\bigvee \text{avg}}$ do not change much when the spatial regularization term is diminished, compared to the other approaches which produce segmentations with smaller regions. This can be better understood looking at the reconstruction quality measures featured by figure 2.16, where these two pruning criteria stabilize around segmentations with approximately 50 regions. It means that the λ value should be severely decreased to obtain more over-segmented partitions.

2.6.2 Cuprite data set

2.6.2.1 Reconstruction errors

Figures 2.18, 2.19 and 2.20 show the quantitative reconstruction quality measures of the different pruning strategies applied over the BPT representations H_μ (mean spectrum region model), H_e (proposed spectral-based region model) and $H_{e\bar{\phi}}$ (proposed spectral-spatial region model) of the Cuprite scene, respectively. The obtained quantitative results present similar trends to the ones obtained for the Pavia University scene. The main difference is that the energy function $\mathcal{E}_\lambda^{\bigvee \text{avg}}$ (2.35) pruning criterion is doing worse than the other proposed pruning criteria. A possible explanation can be formulated when looking at the corresponding optimal partitions (figures 2.21d and 2.21i). In both cases, the partition is composed of a large and under-segmented region (comprising approximately two thirds of the image) and lots of small regions at the center of the image. The over-segmented area corresponds to the mining district, where spectral variability due to minerals is known to happen. The energy $\mathcal{E}_\lambda^{\bigvee \text{avg}}$ admits an optimal partition by minimizing the upper bound on the average RMSE of the regions while maximizing the lower bound on the region size. Having a lot of small regions (thus a large penalty term for each of them) means that this is the configuration which yields the smallest region-wise average RMSE, or alternatively, that it is more costly (in terms of energy) to segment this area with larger regions, thus higher average RMSE values due to the spectral variability. However, using $\mathcal{E}_\lambda^{\bigvee \text{max}}$, which binds the region-wise maximum RMSE instead of the average one, does not lead to the same conclusion as the obtained quantitative values outperform $\mathcal{E}_\lambda^{\bigvee \text{avg}}$,$\mathcal{E}_\lambda^{\sum \text{SID}}$ as well as the conventional height-based pruning approach. An arguable explanation comes again from the analysis of figure 2.21. As a matter of fact, one can see that the region which was strongly under-segmented using $\mathcal{E}_\lambda^{\bigvee \text{avg}}$ is now split into several regions. In such case, it means that the very large region of figures 2.21d and 2.21i has a relatively low average RMSE (which is the reason why it belongs to the optimal cut of $\mathcal{E}_\lambda^{\bigvee \text{avg}}$) but a high maximum RMSE and is then more strongly penalized using $\mathcal{E}_\lambda^{\bigvee \text{max}}$ and is thus forced to split up. This conclusion is supported by the fact that this area of the Cuprite image is more or less segmented the same way by $\mathcal{E}_\lambda^{\sum \text{max}}$ which is also based on the maximal RMSE value of each region.

2.6.2.2 Segmentation results

Figure 2.21 shows the optimal cuts with respect to all five energies, namely (from left to right) $\mathcal{E}_\lambda^{\sum \text{avg}}$, $\mathcal{E}_\lambda^{\sum \text{max}}$, $\mathcal{E}_\lambda^{\bigvee \text{max}}$, $\mathcal{E}_\lambda^{\bigvee \text{avg}}$ and $\mathcal{E}_\lambda^{\sum \text{SID}}$, for the spectral-spatial BPT representation $H_{e\bar{\phi}}$ of the Cuprite scene. As with Pavia University scene, top row shows the optimal partitions with (or close to) 50 regions, while the bottom row shows the optimal partitions with (or close to) 100 regions. Being a scene of a natural landscape, it is difficult to appreciate if the regions are spatially meaningful or not. As already discussed in the case of Pavia, the energy $\mathcal{E}_\lambda^{\sum \text{SID}}$ is the one producing the optimal cuts which seem to better segment all visually salient

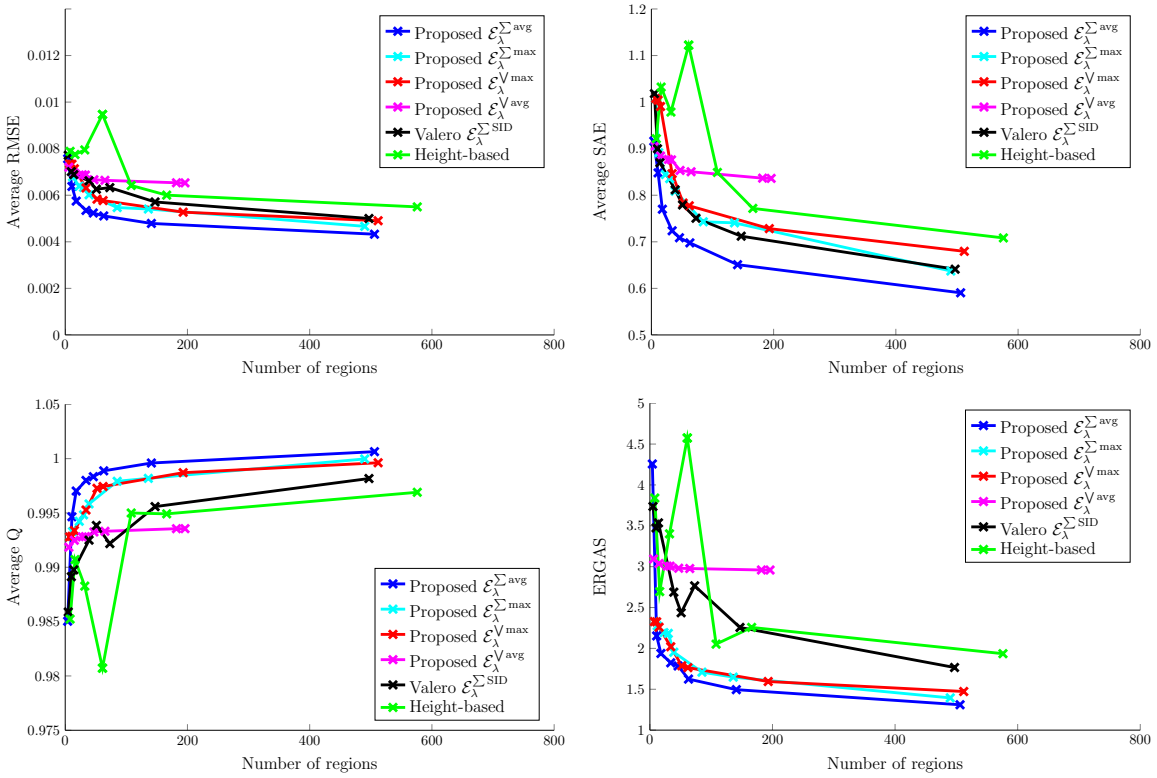


Figure 2.18: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation H_μ (mean spectrum region model) of Cuprite image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

regions. Energies $\mathcal{E}_\lambda^{\sum\text{avg}}, \mathcal{E}_\lambda^{\sum\text{max}}$ and $\mathcal{E}_\lambda^{\bigvee\text{max}}$ yields similar partitions. Onely $\mathcal{E}_\lambda^{\bigvee\text{avg}}$ leads to significantly different results, for the reasons discussed above.

2.7 Conclusion

This chapter has been devoted to the study of the inherent spectral-spatial multimodality of hyperspectral images, and how it can be integrated within hierarchical representations of such images. In particular, the pursued goal was to propose a final segmentation, optimal with respect to the spectral unmixing reconstruction error.

To that purpose, we interest ourselves to the notion of optimality in segmentation. We saw that this framework requires the definition of some energy function which rates how "good" is a given segmentation with respect to the underlying application. Provided this energy function, we observed through several examples (such as the Mumford-Shah functional as well as Markov Random Fields) that finding the partition minimizing the energy is not straightforward, the major challenge being the cardinality and the unstructured nature of the space of partitions

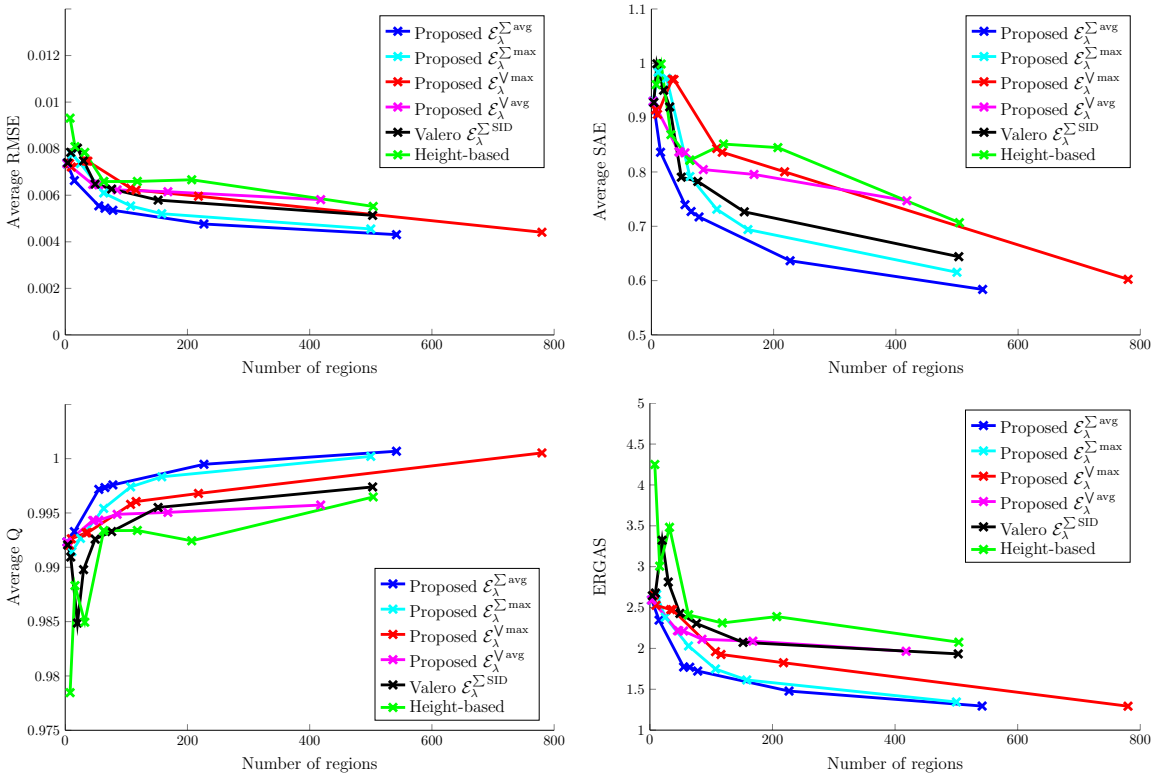


Figure 2.19: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation H_e (proposed spectral-based region model) of Cuprite image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

Π_E. Therefore, conducting the minimization over some constrained space of partitions being the set of all cuts of a hierarchy, came as a natural solution. We therefore reviewed some aspects of hierarchical energy minimization which were first strictly formalized in the work of Guigues [86, 87].

Armed with the fundamental theoretical results being that, under some mild conditions on the definition of the energy function, the optimal cut of the hierarchy can be found by solving Bellman's dynamic program and, when the energy involve some trade-off parameter λ, the optimal cut for different trade-off values can be ordered by refinement, we have developed a new strategy for the representation of hyperspectral images using binary partition trees and concepts from spectral unmixing, in order to finally obtain an optimal segmentation in terms of spectral unmixing reconstruction error. This led us to use the spectral and spatial information bore by hyperspectral images in a synergistic fashion at both steps of the proposed methodology:

– Spectral-spatial information has been incorporated during the construction of the BPT. To that purpose, we proposed two new region models based on the unmixing information. The first one was defined as the set of endmembers induced over each region of the BPT,

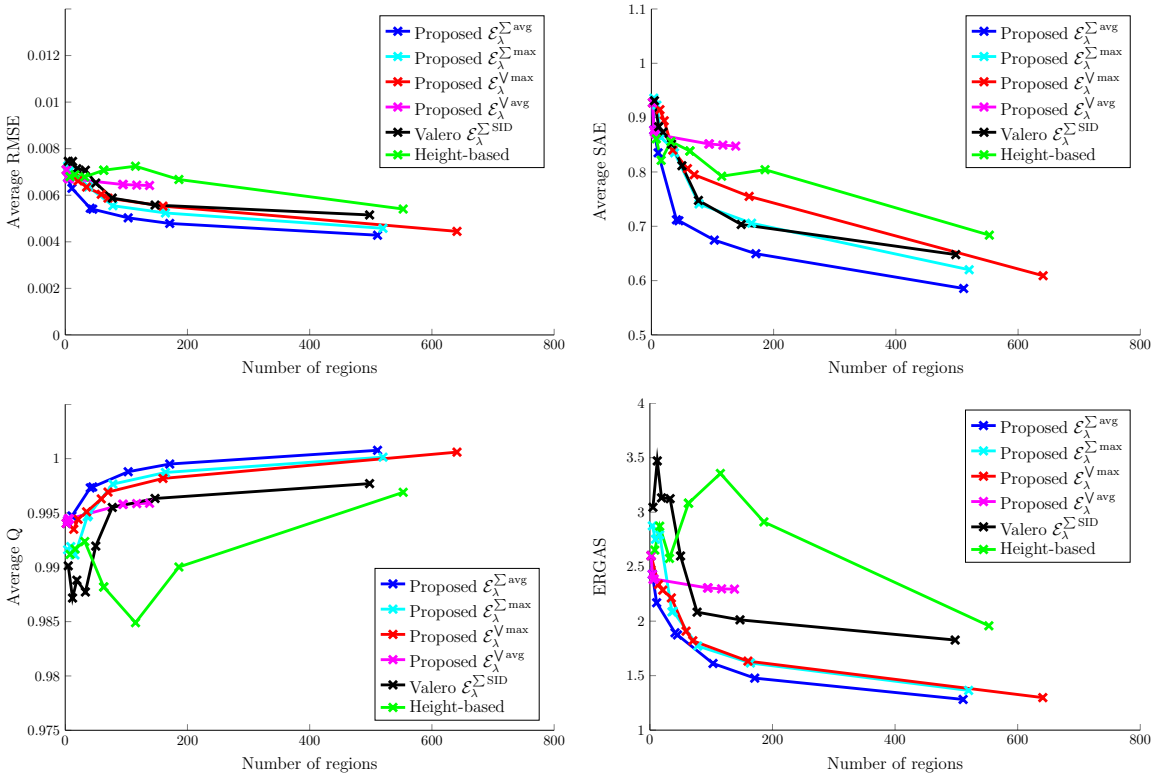


Figure 2.20: Comparison of the different pruning strategies in terms of unmixing reconstruction quality for the BPT representation $H_{e\bar{\phi}}$ (proposed spectral-spatial based region model) of Cuprite image: (top-left) Average RMSE, (top-right) Average SAE, (bottom-left) Average Q and (bottom-right) ERGAS.

the second one also integrating their corresponding abundances. Associated merging criteria were also proposed. To the best of our knowledge, this is the first time in the literature that unmixing information is incorporated in the construction of a BPT.

– We proposed four novel unmixing-based energy functions, defined so their optimal cuts achieve a trade-off between a good spectral reconstruction error, with respect to the unmixing operation, and spatial simplicity. The first two proposed energy functions were formulated as particular instances of a wider class of energy functions, namely *affine separable energies*. Studied by Guigues, there are clear guidelines on the requirements that must be met by such energies to ensure an easy minimization over hierarchies of partitions, and we based the definition of our novel energy functions on those guidelines. For the other two proposed energy functions however, we departed from the framework of affine separable energy functions and investigated what we termed *max-composed energies*. Adapting the work developed by Guigues, and also relying on some more general properties, as drawn by Kiran [101] (exposed in chapter 4), we proved that, under some similar assumptions, all results holding for separable energies were still valid for max-composed energies. These results allowed us to proceed to the minimization of all four proposed unmixing-based energy function in the same way.

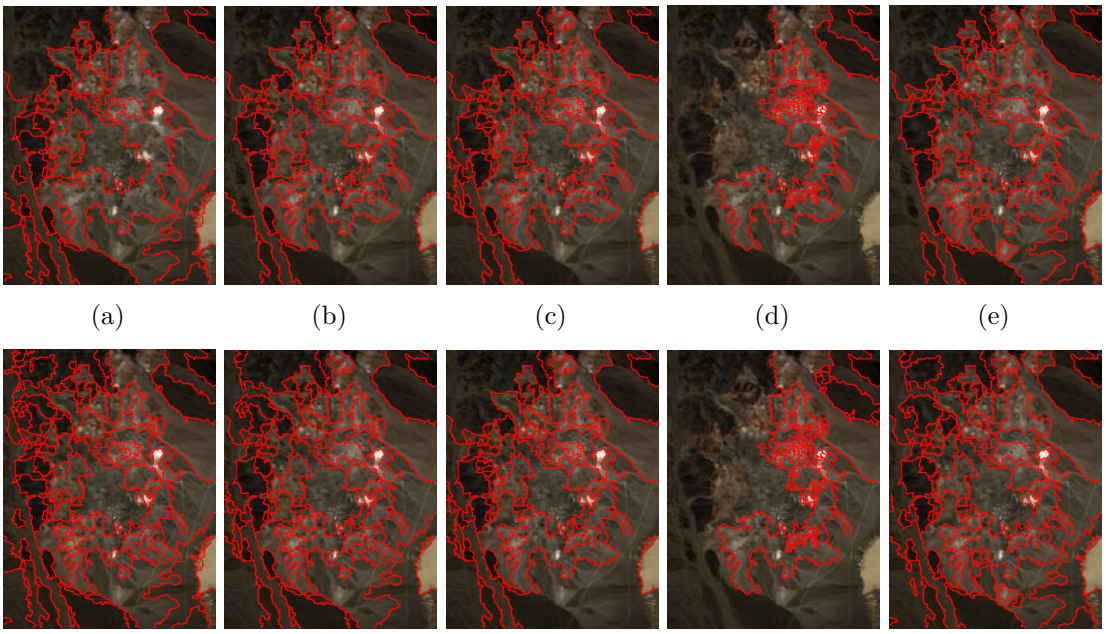


Figure 2.21: Optimal cuts extracted from the BPT representation $H_{e\bar{\phi}}$ of Cuprite scene by minimizing: (a)(f) $\mathcal{E}_\lambda^{\sum\,\text{avg}}$, (b)(g) $\mathcal{E}_\lambda^{\sum\,\max}$, (c)(h) $\mathcal{E}_\lambda^{\bigvee\,\max}$, (d)(i) $\mathcal{E}_\lambda^{\bigvee\,\text{avg}}$, and (e)(j) $\mathcal{E}_\lambda^{\sum\,\text{SID}}$. Top row segmentations have around 50 regions, bottom row segmentations have around 100 regions.

The presented strategy has then been evaluated using reference hyperspectral scenes representing two contexts, urban areas and natural landscapes, at different spatial and spectral resolutions. We compared the segmentation obtained by our proposed energy functions against a classical BPT pruning technique being the height-based cut and against a state-of-the-art energy formulation, proposed by Valero in [204] to achieve hyperspectral image segmentation. The four proposed unmixing-based pruning criteria yielded to segmentations that outperformed the latter two approaches in terms of reconstruction quality. In general, the use of information coming from the unmixing process either in the construction of the BPT representation, by means of the spectral and spectral-spatial region models and merging criteria, or in the pruning of the BPT, by means of the four proposed unmixing-based pruning criteria, showed to have a clear positive impact in the quality of the obtained segmentations.

Although the proposed method has been shown to be a relevant new framework for hyperspectral data interpretation, there are some aspects that may present challenges over time and which deserve a more extensive evaluation. Among them, we list the possibility to use other unmixing-based fitting functions in the definition of the pruning criterion or the evaluation using additional hyperspectral scenes. The proposed approach was found to be useful not only to perform segmentation by taking into account the sub-pixel nature of mixed pixels, but also to perform spectral unmixing using a local-to-global approach in which the optimization criteria is based on the minimization of reconstruction errors at a local scale,

which results in an overall minimization of reconstruction errors that is highly appealing for spectral unmixing applications. As a matter of fact, the natural extension of our approach, proposing an optimal segmentation with respect to the unmixing application, is the processing of the induced endmembers and corresponding abundances over the regions of the optimal partition, in order to provide at the end of the day a similar result to the one furnished by more traditional unmixing approaches being some "global" endmembers and abundance maps for the whole image. A first attempt to such processing has been recently published in [66], where the endmembers induced over all regions of the optimal partition have been stacked together in some sort of library. Following, the abundance maps of the whole image have been obtained by using those endmembers when sparsity is imposed. This preliminary strategy has nevertheless outperformed the traditional global spectral unmixing approach, and has encouraged us to pursue our efforts in this line of research.

Chapter 3

Temporal multimodality

Contents

In this chapter, we now focus on the temporal multimodality, *i.e.*, when several images of a given scene are acquired at different dates. In the most general case, those images may not be acquired with the same sensor, therefore increasing even more the diversity of the resulting multimodal data. However, we only consider the most common case in this chapter, being a single sensor producing images at different acquisition times. In particular, we focus in the following on hyperspectral video sequences. Thanks to the progress made in sensor designing,

it is now possible to acquired sequences of hyperspectral images at near real-time frame rates. However, the extension of traditional video processing techniques from the mature field of computer vision to hyperspectral imagery still faces several challenges due to the very high dimensionality of the resulting data as well as the induced computational burden. In addition, the lack of benchmark hyperspectral video data also make the experimental validation of any new algorithm an issue. On the other hand, the spectral, spatial and temporal information contained in such hyperspectral sequences should lead to the design of robust algorithms, provided that this wealth of information is fully exploited. In this chapter, we propose a novel method to perform object tracking in hyperspectral videos sequences. The tracking is tackled as a sequential object detection process, this latter being performed on a hierarchical decomposition of each frame of the sequence in order to restrain the set of potential candidates for the tracked object. The proposed method is validated in the scenario of chemical gas plume detection and tracking. The present chapter is organized as follows: section 3.1 introduces the temporal multimodality, both for traditional video and hyperspectral images. Section 3.2 reviews the state of the art related to object detection supported by hierarchical decompositions. Then, section 3.3 describes the proposed two-steps method to perform hyperspectral object tracking. Sections 3.4 and 3.5 feature an introduction to hyperspectral chemical gas plume tracking, and the application of the proposed methodology to this problematic, respectively. Results are presented and discussed in section 3.6, and the conclusion is finally drawn in section 3.7.

Materials presented in this chapter have been developed in collaboration with the Department of Mathematics of the University of California, Los Angeles (USA), supported by the National Science Foundation under grant no. DMS-1118971 and no. DMS-0914856. A preliminary version has been published in [196]. The present work has been submitted in a journal version and is currently under review [195].

3.1 Temporal multimodality

3.1.1 Introduction

Temporal multimodality arises when several images of a scene are acquired at different time spots. According to the definition 1.2 of multimodal signals provided in chapter 1, such temporal multimodal data can be formulated as

$$\mathcal{I} = \left\{ \mathcal{I}^{t_1}, \mathcal{I}^{t_2}, \dots, \mathcal{I}^{t_i}, \dots \right\} \tag{3.1}$$

where each modality $\mathcal{I}^{t_i} : E_i \to V_i$ is a particular instance of the scene acquired at time t_i. While nothing forces all the images $\mathcal{I}^{t_i}$ to be produced by the same imaging sensor, this is however the most classical case, to which we will restrict the scope of this chapter. In that situation, one can talk of *multitemporal* data, or simply *video sequence* for $\mathcal{I}$, and each individual modality $\mathcal{I}^{t_i}$ can be referred as a *frame*. All frames share the same spatial support and space of values, $E \equiv E_i, V \equiv V_i \ \forall i$, and the gap of time between two consecutive

acquisitions $t_i - t_{i-1}$ is called the *frame rate* (which is assumed to be constant for the sake of simplicity).

The comparison between several instances of the same image reveals its changes over time. The varying information is due to the complementarity of the multitemporal data, while the remaining, unchanged part, constitutes its redundancy. The interest of such multimodality is rather clear: using the complementarity between consecutive frames is useful to analyze which part of the image are changing, and in which fashion. On the other way around, the redundant part of the information can be used to enhance the robustness of algorithms. As a matter of fact, both sides have been thoroughly investigated in the field of computer vision for traditional gray-scale and color image sequences. The analysis of motion within video sequences has found several applications, such as object tracking [230] or motion estimation and compensation for video compression [111, 235]. Conversely, using the redundancy inherent to video images has been used for instance for patch-based denoising applications, where the similarity between a patch in a given frame and those of a spatial and temporal neighborhood is computed in order to restore the information corrupted by noise [34, 57].

Most commercially available video cameras produce between 25 and 50 images per second. In other words, they have an acquisition frame rate between 25 Hz and 50 Hz (although it is possible to find higher rates of acquisition). Working with hyperspectral sensors however, multitemporal data suffer from a huge decrease in frame rate, as such data is very often provided by airborne or spaceborne sensors. Due to operative constraints, such as the high cost of airborne acquisition campaigns or the revolution time needed to a satellite to stand twice at nadir of the exact same spot of the Earth surface, the average time lapse between two consecutive acquisitions as often been expressed in days. Multitemporal hyperspectral data has therefore been investigated historically in remote sensing mainly for the monitoring of changes occurring over long time periods, due to natural phenomena (such as forestry and environment monitoring [160]) or due to natural disasters, such as floods or volcanic eruptions [201] for instance. To that purpose, a large number of studies have been devoted to hyperspectral change detection, such as statistically-based [32, 33] or kernel-based [39, 40] methods to cite a few.

Thanks to the fast development of imaging sensors, it is now possible to acquire sequences of hyperspectral images at near real-time rates with sensor devices easily operable on the ground by human operators. The combination of the high spectral resolution proper to hyperspectral images with the ability of video sequences to record phenomena evolving with time is appealing for the time monitoring of transient phenomena based on their spatial and spectral properties. However, some additional efforts are required to extend traditional video processing techniques to the high dimensional space structured by hyperspectral data, or to adapt classical hyperspectral processings to multitemporal data whose acquisition frame rate is now in the order of the second. In addition, available benchmark hyperspectral video data-sets are scarce and the lack of ground truth data makes the quantitative evaluation of any novel method very challenging.

3.1.2 Objectives of this chapter

In this chapter, we focus on object tracking in hyperspectral video sequences. Object tracking can be defined as the process of following the motion of points or regions of interest as they evolve with time within a video sequence. Object tracking finds numerous applications in everyday life, such as automated surveillance, motion-based recognition, visual servoing or traffic monitoring, and has been widely studied in the area of computer vision [202, 230] within the framework of traditional video sequences. However, most existing algorithms poorly adapt to the high dimensionality inherent to hyperspectral data. To the best of our knowledge, the only existing tracking method specifically designed and evaluated on real-time hyperspectral video sequences is the one introduced in [15, 213]. It makes use of the mean shift tracker algorithm [53], and the tracked object is represented as a fixed primitive geometric shape and does not adapt well to applications where either the tracked object is non-rigid or where the precise shape of the object is required. The development of new algorithms able to face these challenges is necessary for many real life applications.

Chemical gas plume tracking is a typical application that would surely benefit from the design of such new hyperspectral object tracking methods. As a matter of fact, such application is of great interest for several domains. In the environmental protection field for example, gas plume tracking could be exploited to monitor pollutant gas clouds emitted by industrial sources [233], in order to minimize their impact on the environment and the potential harm they could cause on human population living nearby. In the defense and security area, a possible usage of such tracking method could be to detect the use of chemical gas weapons [71]. Most gases do not respond in the visible spectrum range, but only in a restrained portion of the long-wave infrared (LWIR) domain, hence the need of a fine sampling of the electromagnetic spectrum and the incapacity of classical video techniques to detect (and, *a fortiori*, to track) them. Additionally, a gas plume is a non-rigid object whose shape evolves unpredictably with time. The necessity of a fine spectral description of the scene over time makes hyperspectral video sequences the most suited tool for such detection and tracking application.

In the following, we propose a novel algorithm for hyperspectral object tracking. The method, based on a hierarchical analysis of the frames of the hyperspectral sequence, is able to track a region of interest whose shape may evolve with time, without any prior knowledge about the materials constituting the region. The proposed work, sketched in [196], is based on the sole general assumption that only the object of interest is in motion over a fixed background in the hyperspectral video sequence. It then uses spectral, spatial and temporal information derived from the sequence to perform a sequential object detection process over the hierarchical decomposition of each frame, finally producing the shape and extent of the tracked object. The method is investigated on the scenario of hyperspectral chemical gas plume tracking, and its performances are compared against two state-of-the-art methods for two different data sets.

(a) (b)

Figure 3.1: Example of object detection where (a) only the position and approximate shape and (b) the precise shape of the object (*i.e.*, the bridge) is output.

3.2 Hierarchical object detection

3.2.1 Classical object detection

Object detection is a computer vision application which aims at recognizing and extracting some object of interest from a given image. In other word, the goal of an object detection process is to answer the question: *Is the object of interest present in this given image?* As for most computer vision applications, mimicking the recognition process our brain naturally does turns out to be a real challenge, and several object detection methods have stemmed in the literature. They can be classified in three categories according to the level of details of the detection they produce:

1. The binary output, being the coarsest level, only states whether the object of interest is present in the image or not.
2. The position output, where the location of the object of interest is marked by a simple primitive shape (such as a dot, or a fitting rectangle), as illustrated by figure 3.1a.
3. The position and shape output, which detects the location of the object as well as its precise shape in the image, as in figure 3.1b.

The majority of object detection processes are based on the assumptions that the object of interest is only local with respect to the whole image and can be discriminated for the background using a set characteristic features (such as shape, color homogeneity or texture for instance). Then, due to the locality of the object of interest, the image can be divided into patches, which are subsequently examined to determine whether they contain the object or not by evaluating the presence or absence of the reference features. Regarding the definition of the patches, sliding window approaches have shown to be effective for face [221] or pedestrians [58] detection, as well as recognition of front/side views of cars [178]. The main explanation for this efficiency is that all the sought objects can be roughly approximated by rectangles and

therefore well fit within rectangular windows. Contrarily, sliding windows are not robust to the detection of non rectangular objects. In addition, these approaches suffer from the necessity of fixing the window size (although it can be relaxed by investigating several sizes, but at the cost of a greater computational burden).

A possible solution to alleviate the issues related to sliding windows approaches is the use of a segmentation map, where the spatial support for the sought object would be provided by the various regions constituting the segmentation. In order to further improve this idea, [129, 171] proposed to use several segmentations of the image at various description scales in order to increase the robustness of the resulting so-called *soup of segments* to the detection of objects with various sizes and shapes. This approach is for instance investigated in [3] for the detection of buildings in urban hyperspectral images: a first set of regions are defined as the connected components generated using the morphological profiles of all bands of the image. Meaningful regions are further defined as those with a high spectral homogeneity and large enough size, leading to a batch of potential candidates (this so-called *soup of segments*) for the buildings to retrieve. Finally, the object detection process is conducted by comparing for each candidate region its feature distribution against the feature distribution of the given object of interest, in terms of Kullback-Leibler divergence. Are declared objects all regions whose distance is less than a predefined threshold.

More generally, the object detection process can be formalized as follows: given a set of reference features corresponding to the object of interest $\Omega^{\text{ref}} = \{\omega_i^{\text{ref}}\}$, where each ω_i^{ref} is an individual feature, and given a soup of segments $\mathcal{SS} = \{\mathcal{R} \subseteq E\}$, the object detection process retrieves for each region $\mathcal{R} \in \mathcal{SS}$ its set of features $\Omega^{\mathcal{R}} = \{\omega_i^{\mathcal{R}}\}$ in the image. Following, It evaluates the similarity between Ω^{ref} and $\Omega^{\mathcal{R}}$ for some user chosen distance function $d\left(\Omega^{\text{ref}}, \Omega^{\mathcal{R}}\right)$, defined according to the application. The selected region from the soup of segments is the one minimizing this distance function. Alternatively, all regions below some threshold can be retained.

3.2.2 Hierarchical object detection

Hierarchical image representations are suitable candidates to provide the soup of segments. As a matter of fact, such representations aim at decomposing the image into a set of relevant regions across the image support and at various scales. In addition to naturally providing a finite number of candidate regions (supported by the node of the tree structure), the candidates they propose already bear some meaning (at least with respect to the criterion which was adopted to perform the decomposition). Figure 3.2 shows an example of object detection process conducted over the hierarchical image decomposition presented by figure 1.20. The object to detect is displayed by the leftmost image, the soup of segment is constituted by all 9 regions defining the hierarchy (represented by the tree in the middle), and the retrieved object, extracted from the hierarchy, is depicted by the image on the right. Relevant features to detect the object of interest for this particular example could be the size and orientation of the smallest bounding box enclosing the region as well as its mean color.

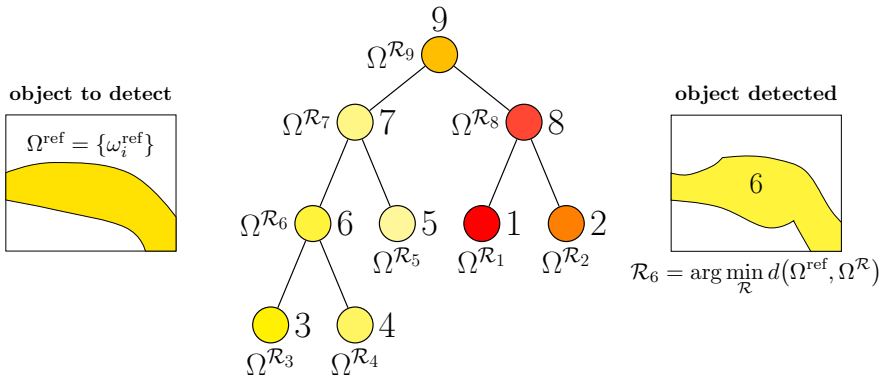


Figure 3.2: Example of hierarchical object detection using the BPT structure of figure 1.20 (page 33).

There are several instances of object detection methods supported by hierarchical decompositions in the literature. The tree of shapes (ToS), also called inclusion tree, has been used by some authors under the assumption that the sought objects of interest can be accurately defined by their level lines. In [155] for instance, the ToS representation is used to extract objects based on their compactness and contrast. Alternatively, [228] proposes to find meaningful objects by identifying the significant local minima of some context-based defined energy along the branches of the ToS. However, as the ToS can be viewed as the merging of a min-tree and a max-tree, its construction remains fully determined by the local extrema of the image. Therefore, its nodes may not always coincide with the objects of interest. In figure 3.1 for instance, the bridge has a sunlit and a shaded part, and the latter one is likely to be confounded with the adjacent buildings having a similar low brightness.

Due to their great flexibility when it comes to their construction, binary partition tree representations have also been investigated as hierarchical supports for object detection. In [175, 218] for instance, BPT representation are used to extract objects with relatively simple shapes (such as faces or road signs) from images by comparing the shape of each region in the BPT against some reference shape models. Still based on BPT representations, [208–210] perform road and building detection in hyperspectral images acquired over urban environments, by making use of spatial features such as the area of the region and of the smallest oriented bounding box containing it, and spectral features such as the correlation between the region mean spectrum and a reference spectrum (asphalt for roads for instance) and some class membership homogeneity. In the following, we will be using the BPT representation to perform the hierarchical analysis of each frame of the hyperspectral video sequence.

3.3 Proposed hyperspectral object tracking method

3.3.1 Generalities on object tracking

Object tracking is the process of following the motion of an object of interest, as it evolves with time in a video sequence. The combination of the always growing computing capacity of

computers and increasing availability of high quality and inexpensive video cameras, along with the need for automated video analysis techniques, have made object tracking a widely studied field of computer vision. Motion-based recognition, automated security and surveillance, video compression and indexing, human-computer interaction, traffic control and monitoring or augmented reality are among the potential applications of object tracking.

Object tracking algorithms are generally organized in two steps that are sequentially addressed:

- The *motion prediction step*, whose goal is to estimate the position of the object in the next frame. This is usually conducted through an interpolation from the current position with the estimations of the motion direction and velocity (plus some margin of error). Motion prediction allows to reduce the search space by defining an area where the object can be found with a high probability.
- The *matching step*, which searches the object in the area predicted by the motion estimation step. It typically involves the definition of reference features for the sought object and their comparison with features derived from candidate objects located in the search space. The tracked object is declared to be the candidate whose features are the closest from the reference ones.

The formulations of the motion prediction and matching steps greatly depend on the representation of the tracked object (punctual objects, primitive shapes such as rectangles or ellipses, exact contours or whole regions for instance) as well as the chosen features to identify the object (histograms, edges or other key points, textures and so on). These choices are in addition motivated by the application scenario and the assumptions which can be made on the object to track. For instance, the point representation is suitable to track an object that appears as almost punctual in the sequence (such as the tracking of a table-tennis ball) but would not be appropriate for complex and non rigid object tracking. The reader is referred to [202, 230] for complete and extensive reviews about classical object tracking algorithms.

Hyperspectral object tracking, on the other hand, remains challenging as the combination of the increased resolution (hence a high dimensionality) of the data and the scarceness of benchmark hyperspectral sequences acquired at real-time rates make the adaptation of classical object tracking algorithms and their validation an issue. To the best of our knowledge, the only tracking method validated on real-time hyperspectral sequences is the one introduced in [15, 213] as an adaptation of the mean shift tracker [53]. The target object, represented as a primitive rectangular shape, has its spectral probability density function computed as a set of N histograms (N being the number of spectral bands for each frame of the hyperspectral sequence). Then, starting from the interpolated position of the object in the new frame (based on the previous positions and velocities), the mean shift tracker iteratively searches for the most probable position of the object by measuring the similarity between the reference spectral probability density function with the one of each candidate object. It is worth noting that, to cope with the high dimensionality of the hyperspectral sequence, a dimensionality reduction step is firstly performed on each frame using random projections [2]. The method is validated for pedestrian and face tracking. However, as for all methods involving a primitive shape representation, it is only able to track the object position and does not retrieve its full shape.

Consequently, we propose in the following a method to perform hyperspectral object tracking which captures both the position and full shape of the object. The tracking is formulated as a sequential object detection procedure and tackled with a hierarchical decomposition of each frame of the hyperspectral sequence, using spectral, spatial and temporal features, by means of a BPT representation. Like classical object tracking algorithms, the proposed methodology is decomposed in a motion prediction step and a matching step, whose descriptions are the matter of the following sections 3.3.2 and 3.3.3. The only basic assumption holding on the video sequence is that only the object of interest is in motion over a fixed background. While it may seem somewhat restrictive, most spectrometer sensors providing hyperspectral video sequences are, for now, still sensors mounted on a tripod, therefore producing sequences of still images with a fixed background.

3.3.2 Motion prediction step

In the following, a region $\mathcal{R}$ is equally handled either as a set ($\mathcal{R} \subseteq E$) or through its indicator function $\mathbb{1}_{\mathcal{R}} : E \to \{0, 1\}$ with $\mathbb{1}_{\mathcal{R}}(x_i) = 1$ if $x_i \in \mathcal{R}$ and 0 otherwise, which leads to a binary visualization of $\mathcal{R}$.

The purpose of the motion prediction step within an object tracking process is to restrict, for each frame $\mathcal{I}^t$, the search space only to a neighborhood where the object is assumed to be found with a high probability. Here we propose to go even one step further: the object $O_t \subseteq E$ being a region of E, the motion prediction step outputs an estimate region $\hat{O}_t$, such that the shape and position of O_t and $\hat{O}_t$ globally coincide. This estimation is then used to steer the matching step to locate a candidate region that is similar to the estimate region both in terms of position and shape.

The method we propose to perform the motion prediction is decomposed in two inner steps as illustrated by figure 3.3. First, the change mask $C_{t-1,t}$ between two consecutive frames $\mathcal{I}_{t-1}$ and $\mathcal{I}_t$ is estimated. This change mask features areas where significant change occurs between $t-1$ and t due to the motion of the object. In a second step, the change mask is combined with the position of the object estimated at $t-1$, denoted O_{t-1} to produce an estimation of position at t, named $\hat{O}_t$.

3.3.2.1 Derivation of the change mask

The first step of the proposed motion prediction step is to derive a change mask $C_{t-1,t}$ between two consecutive frames $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$, in order to highlight the areas which are significantly changing from one frame to the other. Quantifying the significance of a change between two images (being hyperspectral or not) is often carried out with a statistical test [164], which is the idea developed below.

Recall that we position ourselves in the context of a hyperspectral video sequence where the depicted scene is composed of a still background and a moving object, whose (possibly

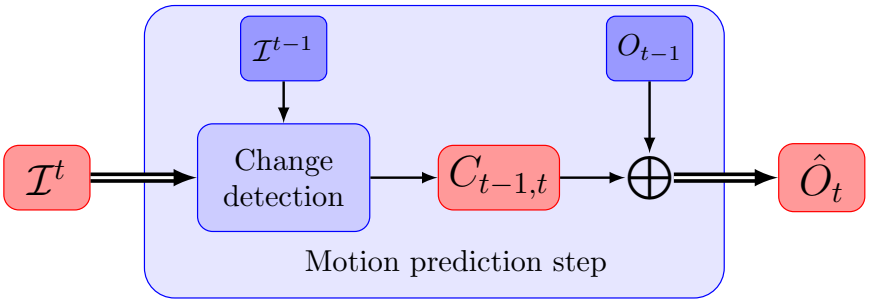


Figure 3.3: Workflow of the proposed motion prediction step, involving first the derivation of a change mask $C_{t-1,t}$ between two consecutive frames $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$ and a binary XOR operation to produce the estimate region $\hat{O}_t$.

unknown) spectral signature, denoted $\mathbf{o}$ is assumed not to vary with time. Following these assumptions, each pixel spectrum $\mathbf{x}_i^t$ belonging to frame $\mathcal{I}^t$ can be expressed as a linear combination of the object signature $\mathbf{o}$ and the background response $\mathbf{b}_i$ at location i (which does not vary with time as the background is supposed to be still), plus some additive noise:

$$\mathbf{x}_i^t = \alpha_i^t \mathbf{o} + (1 - \alpha_i^t)\mathbf{b}_i + \boldsymbol{\eta}^t \tag{3.2}$$

with $\alpha_i^t \in [0, 1]$ being the contribution of the object signature in $\mathbf{x}_i^t$, and $\boldsymbol{\eta}^t$ denotes the noise term. In the framework of the linear mixing model [23], equation (3.2) corresponds to the particular case where each background pixel $\mathbf{b}_i$ and the object signature $\mathbf{o}$ are considered as endmembers. Within this scope, α_i^t can be interpreted as a classical spectral abundance. Similarly, α_i^t can be understood as a model of opacity: $\alpha_i^t = 1$ leads to the case where the object is fully opaque and solely contributes to the spectral signature $\mathbf{x}_i^t$, $\alpha_i^t = 0$ yields the converse interpretation being the absence of object signature in $\mathbf{x}_i^t$ and any other value between 0 and 1 amounts to a measure of opacity.

Following, each pixel signature of the frame difference $\mathcal{I}^{\Delta t} = \mathcal{I}^t - \mathcal{I}^{t-1}$ can then be written

$$\begin{aligned}
\mathbf{x}_i^{\Delta t} &= \mathbf{x}_i^t - \mathbf{x}_i^{t-1} \\
&= \alpha_i^t \mathbf{o} + (1 - \alpha_i^t)\mathbf{b}_i + \boldsymbol{\eta}^t - \left(\alpha_i^{t-1}\mathbf{o} + (1 - \alpha_i^{t-1})\mathbf{b}_i + \boldsymbol{\eta}^{t-1}\right) \\
&= \alpha_i^{\Delta t}(\mathbf{o} - \mathbf{b}_i) + \boldsymbol{\eta}^{\Delta t}
\end{aligned} \tag{3.3}$$

where $\alpha_i^{\Delta t} = \alpha_i^t - \alpha_i^{t-1}$ is the temporal variation of the fractional proportion of object signature in the considered difference pixel $\mathbf{x}_i^{\Delta t}$, and $\boldsymbol{\eta}^{\Delta t} = \boldsymbol{\eta}^t - \boldsymbol{\eta}^{t-1}$. Consequently, a variation in the proportion of the object signature at pixel position i between time instances $t-1$ and t yields $\alpha_i^{\Delta t} \neq 0$, while $\alpha_i^{\Delta t} = 0$ when no change occurs. This observation naturally leads to formulate the following two-hypotheses test:

$\mathbf{H}_0 : \alpha_i^{\Delta t} = 0$, the proportion of object signature $\mathbf{o}$ in x_i does not change between $t-1$ and t,

$\mathbf{H}_1 : \alpha_i^{\Delta t} \neq 0$, the proportion of object signature $\mathbf{o}$ in x_i changes between $t-1$ and t.

Therefore, deriving the change mask $C_{t-1,t}$ between $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$ amounts to testing whether each pixel signature of the frame difference $\mathcal{I}^{\Delta t}$ is more likely to follow hypothesis $\mathbf{H}_0$ or $\mathbf{H}_1$.

Provided that the probability distribution functions of $\mathbf{x}_i^{\Delta t}$ under both hypotheses are known, the Neyman-Pearson lemma states that the Likelihood Ratio Test is the most powerful test given a significance level [176].

In hyperspectral imagery, it is classically assumed that the additive noise follows a Gaussian distribution with zero mean and covariance Σ (see for instance [131, 186]). Therefore, provided that $\boldsymbol{\eta}^{\Delta t} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, it is possible to formulate the probability distribution function f of $\mathbf{x}_i^{\Delta t}$ under both hypotheses $\mathbf{H}_0$ and $\mathbf{H}_1$:

$$\begin{aligned} \mathbf{H}_0 : &\quad f(\mathbf{x}_i^{\Delta t} | \alpha_i^{\Delta t} = 0) \sim \mathcal{N}(\mathbf{0}, \Sigma) \\ \mathbf{H}_1 : &\quad f(\mathbf{x}_i^{\Delta t} | \alpha_i^{\Delta t} \neq 0) \sim \mathcal{N}\left(\boldsymbol{\mu}_i^{\Delta t}, \Sigma\right) \end{aligned} \tag{3.4}$$

with $\boldsymbol{\mu}_i^{\Delta t} = \alpha_i^{\Delta t}(\mathbf{o} - \mathbf{b}_i)$.

Therefore, detecting a change in the time difference frame reduces to testing whether each pixel difference is drawn from a zero-mean Gaussian distribution (with covariance matrix Σ) or not. However as the object spectral signature $\mathbf{o}$ is *a priori* unknown, solving the two-hypotheses test (3.4) involves instead a Generalized Likelihood Ratio Test (GLRT) whose expression $\Lambda(\mathbf{x}_i^{\Delta t})$ for the pixel $\mathbf{x}_i^{\Delta t}$ is the following:

$$\Lambda(\mathbf{x}_i^{\Delta t}) = \frac{\max\limits_{\boldsymbol{\mu}_i^{\Delta t} \neq 0} f(\mathbf{x}_i^{\Delta t} | \mathbf{H}_1)}{f(\mathbf{x}_i^{\Delta t} | \mathbf{H}_0)} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \tag{3.5}$$

The unknown mean $\boldsymbol{\mu}_i^{\Delta t}$ that maximizes $f(\mathbf{x}_i^{\Delta t} | \mathbf{H}_1)$ is known to be the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\mu}}_i^{\Delta t}$ of $\boldsymbol{\mu}_i^{\Delta t}$. Testing the pixel $\mathbf{x}_i^{\Delta t}$ alone would yields $\hat{\boldsymbol{\mu}}_i^{\Delta t} = \mathbf{x}_i^{\Delta t}$ and a test statistic being

$$\Lambda(\mathbf{x}_i^{\Delta t}) = \left(\mathbf{x}_i^{\Delta t}\right)^T \Sigma^{-1} \mathbf{x}_i^{\Delta t} \tag{3.6}$$

However, for the reasons further explained in section 3.3.4, we rather make use also of neighboring pixels of $\mathbf{x}_i^{\Delta t}$, selected in a $S = S_{\text{width}} \times S_{\text{height}}$ window (hereafter set to 3×3) centered on $\mathbf{x}_i^{\Delta t}$. The GLRT expression (3.5) can the be rewritten as

$$\Lambda(\mathbf{x}_i^{\Delta t}) = \frac{\max\limits_{\boldsymbol{\mu}_i^{\Delta t} \neq 0} \prod\limits_{\mathbf{x}_i^{\Delta t} \in S} \left[f(\mathbf{x}_i^{\Delta t} | \mathbf{H}_1) \right]}{\prod\limits_{\mathbf{x}_i^{\Delta t} \in S} \left[f(\mathbf{x}_i^{\Delta t} | \mathbf{H}_0) \right]} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \tag{3.7}$$

Plugging the following MLE of $\hat{\boldsymbol{\mu}}_i^{\Delta t}$,

$$\hat{\boldsymbol{\mu}}_i^{\Delta t} = \frac{1}{S} \sum_{\mathbf{x}_i^{\Delta t} \in S} \mathbf{x}_i^{\Delta t} \tag{3.8}$$

into equation (3.7), and solving for $\mathbf{x}_i^{\Delta t}$ (the computational details are presented in appendix B) finally yields to

$$\Lambda(\mathbf{x}_i^{\Delta t}) = S(\hat{\boldsymbol{\mu}}_i^{\Delta t})^T \Sigma^{-1} \hat{\boldsymbol{\mu}}_i^{\Delta t} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \tag{3.9}$$

Algorithm 2 Proposed change detection procedure.

Require: $\mathcal{I}^t, \mathcal{I}^{t-1}, \Sigma, \gamma_{\text{GLRT}}$

1. $\mathcal{I}^{\Delta t} = \mathcal{I}^t - \mathcal{I}^{t-1}$ ▷ Frame difference.

for all $x_i \in E$ **do** ▷ Test each pixel site x_i.

2. $\hat{\boldsymbol{\mu}}_i^{\Delta t} \leftarrow \text{MLE}\left(\boldsymbol{\mu}_i^{\Delta t}\right)$

3. $\Lambda(\mathbf{x}_i^{\Delta t}) \leftarrow S(\hat{\boldsymbol{\mu}}_i^{\Delta t})^T \Sigma^{-1} \hat{\boldsymbol{\mu}}_i^{\Delta t}$ ▷ S = number of pixels to estimate the MLE $\hat{\boldsymbol{\mu}}_i^{\Delta t}$

if $\Lambda(\mathbf{x}_i^{\Delta t}) \geq \gamma_{\text{GLRT}}$ **then** ▷ $\Lambda(\mathbf{x}_i^{\Delta t}) \sim \mathbf{H}_1$

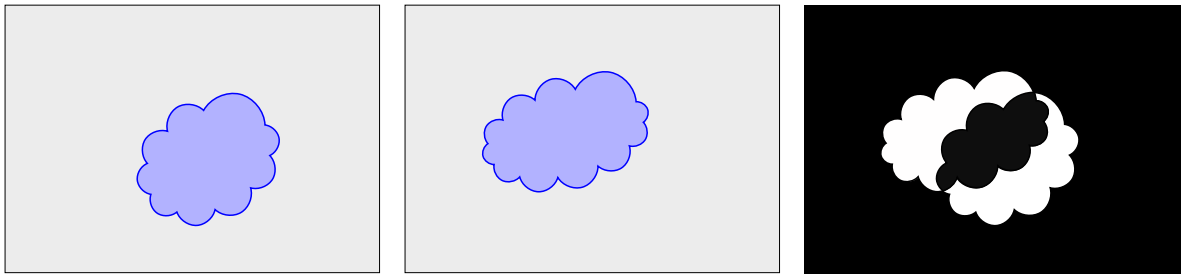
4. $C_{t-1,t}(x_i) \leftarrow$ **true** ▷ There is significant change in x_i between $t-1$ and t.

else ▷ $\Lambda(\mathbf{x}_i^{\Delta t}) \sim \mathbf{H}_0$

5. $C_{t-1,t}(x_i) \leftarrow$ **false** ▷ There is no significant change in x_i between $t-1$ and t.

end if

end for



(a) O_{t-1} (b) O_t (c) $C_{t-1,t}$

Figure 3.4: The object O_t (b) can be retrieve from O_{t-1} (a) and the change mask $C_{t-1,t}$ (c) by $O_t = O_{t-1} \bigoplus C_{t-1,t}$.

The probability distribution of $\Lambda(\mathbf{x}_i^{\Delta t})$ under $\mathbf{H}_0$ and $\mathbf{H}_1$ are given in terms of χ^2_N with N degrees of freedom (being the number of spectral channels in the hyperspectral frame) and non-central $\chi^2_{N,\phi}$ with N degrees of freedom and non-centrality parameter $\phi = \Lambda(\mathbf{x}_i^{\Delta t})$ [186], respectively.

Knowing the distribution of the GLRT under both hypotheses allows one to set γ_{GLRT} to achieve either a predefined probability of false alarm or a probability of detection (discussed in section 3.3.4), and to further state whether some change is occurring in $\mathbf{x}_i$ between frames $t-1$ and t by thresholding accordingly. The binary change mask $C_{t-1,t}$ is finally obtained by performing this threshold operation for all pixels of the frame difference. The proposed change detection process is summarized by algorithm 2.

3.3.2.2 Estimation of the position

Under the assumption of a single moving object overlaying a fixed background, the change mask $C_{t-1,t}$ is composed of two categories of regions:

- Regions being left by the object. Those are made of all pixels x_i which were occupied

by the object in frame $\mathcal{I}^{t-1}$ ($\alpha_i^{t-1} > 0$) but no longer in $\mathcal{I}^t$ ($\alpha_i^t = 0$). They correspond to the white region on the bottom right of figure 3.4c.

- Regions invaded by the object. Contrarily, those regions are composed of all pixels which are reached by the object in frame $\mathcal{I}^t$ (thus, $\alpha_i^{t-1} = 0$ and $\alpha_i^t \geq 0$), and are depicted by the top left white region in figure 3.4c.

Intuitively, the new position of the object, O_t is composed of the previous position of the object, O_{t-1}, minus the regions that have been left by the object plus the regions that have been reached by it, as depicted by figure 3.4. Mathematically, this can be formulated by:

$$\hat{O}_t = O_{t-1} \bigoplus C_{t-1,t} \tag{3.10}$$

where $\bigoplus$ denotes the binary XOR operation. However, as both the previously known position of the object O_{t-1} and change mask $C_{t-1,t}$ may not be fully accurate, equation (3.10) is better to be used as a simple estimate of the new position and shape of the object. This estimate, output of the motion prediction step as shown by figure 3.3 [1], is going to be used as target and further refined during the object detection process of the following matching step.

3.3.3 Matching step

The proposed matching step, illustrated by the work-flow figure 3.5 involves a hierarchical decomposition H_t of the current frame $\mathcal{I}^t$ and a subsequent object detection processed with this decomposition as support. Using a hierarchical decomposition bears several advantages:

- It drastically reduces the search space by representing the frame as a set of hierarchically nested regions. The set of candidate objects is only composed of regions that are supported by a node in the decomposition.
- It ensures to represent the frame at various scales, which is valuable as the size of the tracked object may evolve along the sequence.
- It allows to enjoy from all the efficient tree-based processing techniques already available in the literature.

Here we propose to use the BPT representation to handle the hierarchical decomposition of the frame. This choice is motivated by the great flexibility of the BPT construction due to all possible combinations of region models and associated merging criteria. The selection of these parameters has to be done in accordance with the pursued goal. If appropriately done, it is however very likely to yield a meaningful hierarchical decomposition, in the sense that the tracked object can be identified as a region of this decomposition.

Therefore, the matching process aims at retrieving in the BPT structure the region that represents the tracked object. In order to do so, a set of reference features for the tracked object is defined, and each candidate region has its own set of similarly defined features evaluated against the reference set. The region whose features match the reference the best is declared to correspond to the tracked object.

1. The object is displayed as a cloud in figure 3.4 not to anticipate on the following gas plume tracking application, but rather to symbolize the fact that the tracked object can be non rigid and of irregular shape.

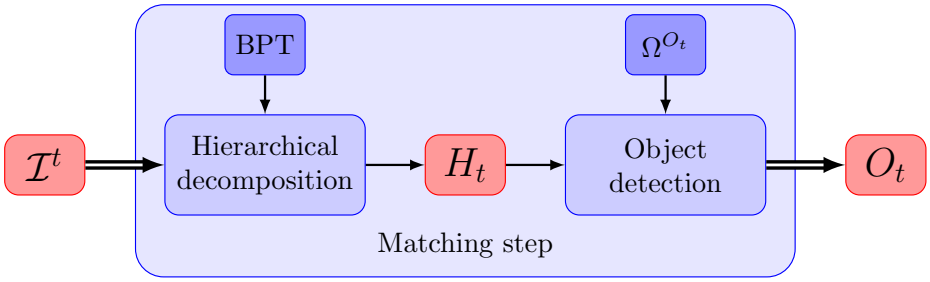


Figure 3.5: Workflow of the proposed matching step, involving first a hierarchical decomposition of the current frame $\mathcal{I}^t$ by means of a BPT representation, and a further object detection procedure to identify the tracked object based on its reference features Ω^{O_t}.

More specifically, let $\Omega^{O_t} = \{\omega_i^{O_t}\}$ be the set of chosen reference features for the tracked object O_t, where $\omega_i^{O_t}$ denotes an individual feature. Using hyperspectral video sequences allows to make the most of spectral, spatial and temporal features to describe the tracked object. For each frame, the object detection procedure is done in three stages:

1. First, each region $\mathcal{R}$ of the BPT H_t has its similarly defined features collected in a set $\Omega^{\mathcal{R}} = \{\omega_i^{\mathcal{R}}\}$.
2. Then, each region $\mathcal{R}$ has its set of features $\Omega^{\mathcal{R}}$ evaluated against the reference Ω^{O_t}. This implies the definition of a similarity measure d_i for each pair of individual features $\omega_i^{\mathcal{R}}$ and $\omega_i^{O_t}$ so the overall matching distance $d(\mathcal{R}, O_t)$ can be formulated as follows:

$$d(\mathcal{R}, O_t) = \sum_{\omega_i \in \Omega} \rho_i d_i \left(\omega_i^{\mathcal{R}}, \omega_i^{O_t}\right) \tag{3.11}$$

   where the ρ_i's are optional weights that can be set to stress the importance of some individual features against others.
3. Finally, the region $\mathcal{R}^\star$ whose features are the closest from the reference ones is retrieved,

$$\mathcal{R}^\star = \underset{\mathcal{R} \in H_t}{\operatorname{argmin}}\, d(\mathcal{R}, O_t)\,. \tag{3.12}$$

This region $\mathcal{R}^\star$ becomes the representation of the object in the current frame $\mathcal{I}^t$, $O_t \equiv \mathcal{R}^\star$, and is going to be used for the motion prediction step in the next frame $\mathcal{I}^{t+1}$.

3.3.4 Initialization of the object tracking procedure

The motion prediction and matching steps developed above in sections 3.3.2 and 3.3.3 are sequentially addressed in order to track the object of interest. In order to trigger the tracking process however, an initial detection needs to be performed to identify the object to track. This is the matter of the initialization phase. In the following, we assume that the frame in which the object starts moving (and thus where the object tracking process must be launched) is unknown. However, we presume that a few (at least two) still frames $\mathcal{I}^{t_1}, \ldots, \mathcal{I}^{t_{N_s}}$

are available prior to the object being in motion. This assumption seems however reasonable in a context of surveillance, where nothing is moving in most of the frames of the sequence.

To determine the point at which some motion appears in the sequence, the change detection procedure described in section 3.3.2.1 is applied for each new incoming frame $\mathcal{I}^t$, and a change mask $C_{t-1,t}$ is generated. If this change mask remains empty (all pixels of the frame difference $\mathcal{I}^t - \mathcal{I}^{t-1}$ have been found not to feature any change), then it is stated that no motion is occurring in the sequence yet. Conversely, if at least one pixel was found to be changing between $t-1$ and t (appearing as a 1 in the change mask $C_{t-1,t}$), then it is assumed that the object has started moving, and the object tracking process is triggered. Deriving the change mask requires the knowledge of the covariance matrix Σ of the noise $\boldsymbol{\eta}^{\Delta t}$, which is unknown in practice. However, due to the assumptions that several (say N_s) still frames are available, one can compute $N_s - 1$ frame differences featuring only instances of $\boldsymbol{\eta}^{\Delta t}$, which can be used to derive the sample covariance matrix $\hat{\Sigma}$ of Σ. In addition, due to the (ideally) large number of samples on which $\hat{\Sigma}$ is computed, it seems fair to state that the sample covariance $\hat{\Sigma}$ is a very good approximation of Σ, which is the reason why this latter was used in the derivation of the change mask instead of $\hat{\Sigma}$.

The tracking process being triggered once at least one pixel has been stated to change between $t-1$ and t, the change detection test needs not to suffer from any false alarm at all, as the tracking would be engaged too early otherwise. Therefore, instead of testing each pixel of the frame difference $\mathbf{x}_i^{\Delta t}$ individually in the change detection process, it is tested along with its 3×3 neighbors. Therefore, the pixel will be marked as changing if and only if some change is also happening in its direct neighborhood, decreasing the risk of false alarms with respect to the individual testing case (or conversely, guaranteeing that all pixels marked as change are really changing). In the derivation of the GLRT, this translates as the MLE $\hat{\boldsymbol{\mu}}_i^{\Delta t}$ being equal to $\frac{1}{S} \sum_{\mathbf{x}_i^{\Delta t} \in S} \mathbf{x}_i^{\Delta t}$ instead of simply $\mathbf{x}_i^{\Delta t}$.

The other main consequence of this no false alarm policy is related to the setting of the γ_{GLRT} threshold. Usually, this threshold is derived using the distribution of the GLRT under hypothesis $\mathbf{H}_0$ in order to achieve a given probability of false alarm p_{FA}. In particular, it requires to invert the cumulative distribution function of the GLRT under $\mathbf{H}_0$. In our case however, we wish to have $p_{FA} = 0$ and thus needs to invert the distribution of the GLRT under $\mathcal{H}_1$ to achieve a given probability of detection p_D. It is known from [186] that the GLRT $\Lambda(\mathbf{x}_i^{\Delta t})$ follows a non-central $\chi^2_{N,\phi}$ distribution under $\mathbf{H}_1$, with N degrees of freedom being the number of spectral bands in the frame, and $\phi = \Lambda(\mathbf{x}_i^{\Delta t})$ being the non-centrality parameter. The threshold γ_{GLRT} and the probability of detection p_D are linked with the following relationship:

$$p_D = \int_{\gamma_{\text{GLRT}}}^{+\infty} \chi^2_{N,\phi}(t)dt \tag{3.13}$$

$$= 1 - \mathbf{X}^2_{N,\phi}(\gamma_{\text{GLRT}}) \tag{3.14}$$

with $\mathbf{X}^2_{N,\phi}$ being the cumulative distribution function of $\chi^2_{N,\phi}$ defined as

$$\mathbf{X}^2_{N,\phi}(t) = \int_{-\infty}^{t} \chi^2_{N,\phi}(z)dz \tag{3.15}$$

Therefore, using (3.14), one can set the value of γ_{GLRT} to achieve a given p_D as

$$\gamma_{\text{GLRT}} = (\mathbf{X}^2_{N,\phi})^{-1}(1 - p_D) \tag{3.16}$$

In practice however, the cumulative distribution function $\mathbf{X}^2_{N,\phi}$ has no close form expression and is computationally slow to invert. However, it was shown [141, pp.22-24] that if the random variable $\mathbf{Y}$ follows a non-central $\chi^2_{N,\phi}$ distribution with N degrees of freedom and non-centrality parameter ϕ, then

$$\frac{\mathbf{Y} - (N + \phi)}{\sqrt{2(N + 2\phi)}} \xrightarrow{p} \mathcal{N}(0, 1) \text{ when } N \to +\infty \text{ or } \phi \to +\infty \tag{3.17}$$

where $\xrightarrow{p}$ denotes the convergence in probability. Here, the number of degrees of freedom N is equal to the number of spectral channel in the hyperspectral frame, which is typically several hundreds and can be considered high enough for the approximation (3.17) to hold. Therefore, the value of γ_{GLRT} can be derived by inverting the cumulative distribution function of a standard normal distribution (provided that the proper shift and scaling described by (3.17) is applied) instead of this of the non-central χ^2 distribution (as prescribed by (3.16)) to achieve a given probability of detection p_D. Note finally that γ_{GLRT} varies from one pixel to the other, as it is linked to the value of $\phi = \Lambda(\mathbf{x}_i^{\Delta t})$.

3.3.5 Summary

Eventually, the proposed hyperspectral object tracking by means of hierarchical decomposition is summarized by algorithm 3.

3.4 Chemical gas plume tracking

From hereon, we specifically focus on the chemical gas plume tracking application. The current section first describes the underlying physical nature of hyperspectral gas plume data set (section 3.4.1) prior to reviewing the related state-of-the-art for the detection and tracking of such plumes (in section 3.4.2) and finally introduces the data sets on which the previously proposed method is investigated (section 3.4.3).

3.4.1 Radiative transfert theory

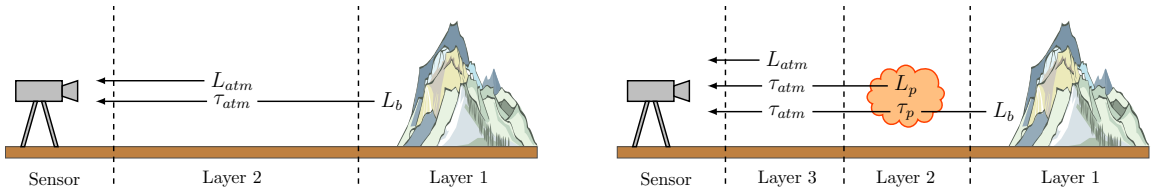
Hyperspectral sensors measure the *radiance*, that is, the amount of electromagnetic energy emitted by the scene. The physical nature of this energy depends on the scanned spectral

Algorithm 3 Proposed hyperspectral object tracking algorithm.

Require: a hyperspectral sequence $\mathcal{I} = \{\mathcal{I}^{t_1}, \ldots, \mathcal{I}^{t}, \ldots\}$

Init. $C_{t-1,t} \leftarrow$ **False**(N_x, N_y)
$O_{t-1} \leftarrow$ **False**(N_x, N_y)
▷ N_x and N_y are the number of rows and columns of each hyperspectral frame, respectively.

for $\mathcal{I}^t \in \mathcal{I}$ **do** ▷ Process each new frame of the sequence.
1. $\mathcal{I}^{\Delta t} \leftarrow \mathcal{I}^t - \mathcal{I}^{t-1}$ ▷ Frame difference.
2. $C_{t-1,t} \leftarrow$ **ChangeDetection**$\left(\mathcal{I}^{\Delta t}, p_D\right)$ ▷ Change detection on $\mathcal{I}^{\Delta t}$.
if IsFalse$(C_{t-1,t})$ **then** ▷ No change between $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$.
3. $t \leftarrow t+1$
Go to step **1**. ▷ Repeat until some change (motion) is detected.
else
4. $\hat{O}_t = O_{t-1} \bigoplus C_{t-1,t}$ ▷ Motion prediction.
5. $H_t \leftarrow$ **BuildBPT**$(\mathcal{I}^t)$ ▷ Hierarchical decomposition of $\mathcal{I}^t$.
6. $O_t \leftarrow$ **ObjectDetection**$\left(H_t, \Omega^{O_t}\right)$ ▷ Object detection.
7. $t \leftarrow t+1$ ▷ Go to next frame.
end if
end for



(a) (b)

Figure 3.6: Illustration of the two layers (a) and three layers (b) models for the at-sensor radiance, respectively.

range. In the LWIR range, this emitted energy is governed by the radiative transfer theory. Therefore, in order to understand the physical nature of the used data sets, it is worth briefly describing the radiative transfer theory.

The radiance of a material is defined as the amount of electromagnetic radiation which passes through or is emitted from a particular unit area per solid angle, and is expressed in $\text{Wsr}^{-1}\text{m}^{-2}$. Consequently, the radiance of a material is expressed over the whole electromagnetic spectrum. However, the sensor cannot capture the radiance over the whole spectrum, but rather at some particular wavelengths. Thus, the *spectral radiance* is defined as the radiance of a material at a given wavelength, and is expressed in terms of $\text{Wsr}^{-1}\text{m}^{-3}$. When no plume is present in the scene, the spectral radiance $L(\lambda)$ (which is a function of the wavelength λ) reaching the sensor can be expressed according to the *two layers model*:

$$L(\lambda) = L_{atm}(\lambda) + \tau_{atm}(\lambda) L_b(\lambda) \tag{3.18}$$

as depicted by figure 3.6a. In such case, the radiance received by the sensor is expressed as the sum of the atmosphere spectral radiance $L_{atm}(\lambda)$ and the background spectral radiance $L_b(\lambda)$ modulated by the atmosphere transmittance $\tau_{atm}(\lambda)$. This latter quantity is defined as the ratio of the amount of light leaving a medium (the atmosphere in this case) with respect to the amount of light entering this medium. The presence of a plume in the scene has two effects: it absorbs part of the radiance emitted by the background due to its own transmittance $\tau_p(\lambda)$ and it emits its own radiations $L_p(\lambda)$. Therefore, equation (3.18) transforms into the so-called *three layers model*, depicted by figure 3.6b and expressed as

$$L(\lambda) = L_{atm}(\lambda) + \tau_{atm}(\lambda)L_p(\lambda) + \tau_{atm}(\lambda)\tau_p(\lambda)L_b(\lambda) \tag{3.19}$$

It is often assumed that the contribution of the atmospheric radiance with respect to the plume and background radiances can be neglected. Also, the atmospheric transmittance can be approximated to 1 when the distance between the release point and the sensor is short (typically no more than a few kilometers), meaning that the atmosphere allows all the signal to pass through unaffected [29]. Under those assumptions, the at-sensor spectral radiance can be written

$$L(\lambda) = L_p(\lambda) + \tau_p(\lambda)L_b(\lambda) \tag{3.20}$$

Up to some constant coefficient, equation (3.20) resembles equation (3.2), which also modeled each pixel signature as a linear combination of the plume and background signatures.

Usually, spectral radiance is converted into spectral *emissivity* prior to any further processing, as the latter plays in the LWIR domain the same role as the reflectance does in the visible domain. As a matter of fact, to each material is associated a unique spectral emissivity in the LWIR domain, which acts as a signature proper to the material [13]. The emissivity of a material, $\epsilon(\lambda)$, is defined as the ratio of the energy radiated by this particular material to the energy radiated by a black body at the same temperature. While the former is the quantity acquired by the sensor, the latter is described using Planck's black body law:

$$B(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{exp\left(\frac{hc}{kT\lambda}\right) - 1} \tag{3.21}$$

where T is the temperature of the surface in Kelvin, h is the Planck's constant, c is the speed of light and k is the Boltzmann's constant. The spectral emissivity of each pixel is retrieved from the spectral radiance through the use of some Temperature Emissivity Separation (TES) algorithm [80], which operates in two steps: first, the apparent temperature of each pixel is estimated by inverting Planck's law. Each pixel spectral radiance is then divided by its estimated black body curve, evaluated at the proper wavelengths, to finally obtain the spectral emissivity.

3.4.2 State of the art

The detection of gas plumes has been already addressed in the literature [130], where most techniques can be categorized either as anomaly/target detectors or as clustering-based methods.

3.4.2.1 Anomaly/target detectors

The most popular and natural approach is to consider the gas plume as anomalous with respect to the background, and thus make use of conventional anomaly or target detectors. More specifically, such methods decide whether some signal of interest (being the anomaly or the target signature) is present or not in each pixel signature $\mathbf{x}$. This involves the solving of the two-hypotheses test:

$$\begin{aligned} \mathbf{H}_0 : &\quad \mathbf{x} \sim f_0 \text{ (target absent)} \\ \mathbf{H}_1 : &\quad \mathbf{x} \sim f_1 \text{ (target present)} \end{aligned} \tag{3.22}$$

where f_0 and f_1 are the assumed distributions of the pixels signatures with and without the presence of the target, respectively.

According to the further hypotheses made on the general statement of (3.22) (the target being either expressed as a full or a mixed pixel, with a fixed or probabilistic signature, the background distribution being either structured or unstructured), solving (3.22) leads to different detectors. Considering for instance a full pixel target yields the Adaptive Matched Filter (AMF), operated in [185]. However, a full pixel assumption implies that the plume is opaque so its signal is not influenced by the background behind it, which is erroneous in practice and rather suggests the use of anomaly/target detectors for mixed pixels. Depending on the variability and structure assumed for the background (a structured background is described by a subspace model, while an unstructed one is characterized by a statistical distribution), one can finally implement the Adaptive Matched Subspace Detector (AMSD), investigated in [29, 30, 147] and described in the following section 3.5.4, the Clutter Matched Filter (CMF) used in [71, 211], the Adaptive Cosine/Coherence Estimator (ACE) or the Orthogonal Subspace Projection (OSP) considered in [132]. Performance comparison of AMF and AMSD for gas plume detection can be found in [109]. For futher details about previous anomaly and target detectors, the reader is referred to [131, 133, 186].

The major drawback of anomaly and target detector methods for gas plume detection is that they cannot be operated without a reference target spectrum, often estimated using spectral libraries [149, 181]. Moreover, they do not use any temporal information since the target detection process is applied on each frame independently.

3.4.2.2 Clustering-based methods

A second popular approach that recently emerged is to address the plume detection as a clustering problem. In that case, it is assumed that properties of the spectral signature of the plume are sufficiently different from those of the background so it is possible to compose a cluster solely containing the plume. Hierarchical clustering is for instance investigated in [91] (note that it also requires a reference target spectrum). Alternately, graph Laplacian-based spectral clustering is notably considered in [79], as well as in [93, 138] and [163] where it is used as an initialization for more powerful clustering methods, such as semi-supervised diffused interface clustering [137] for the former, and robust non-negative matrix factorization [234] for the latter.

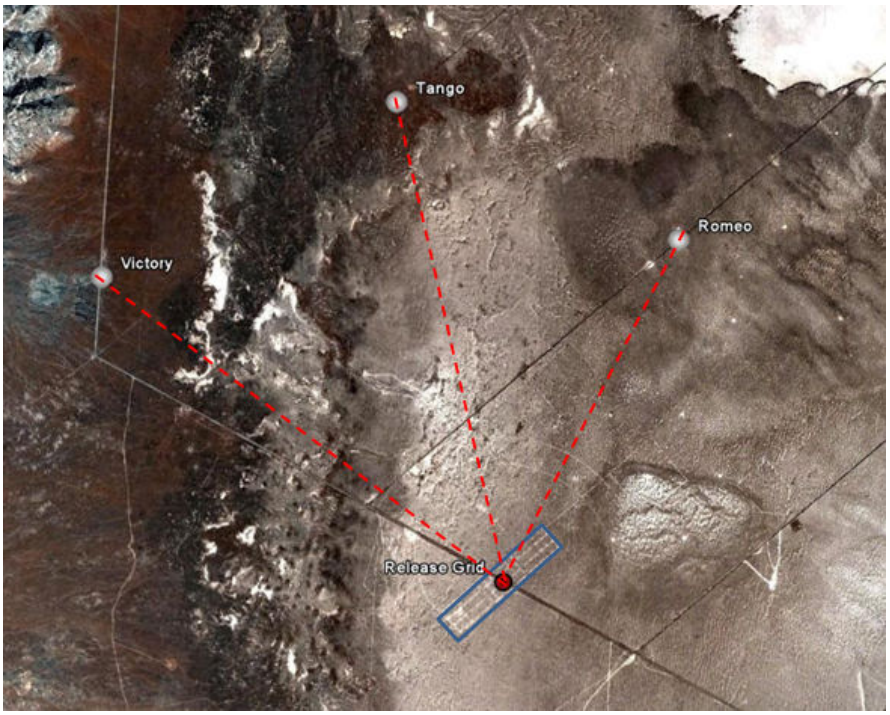


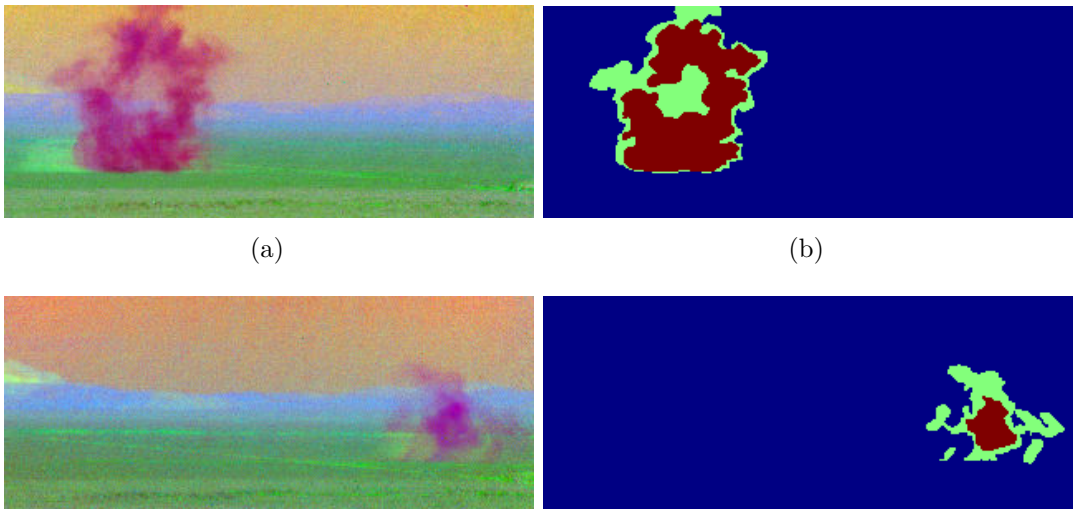
Figure 3.7: Location of the three FIRST sensors recording the chemical plume release (image from [29]).

As clustering methods aims at dividing each frame of the sequence into clusters, they do not take temporal information into account innately. To tackle this issue, several consecutive frames are stacked together prior to the clustering operation in [93, 138, 163] in order to introduce some temporal coherency. The drawback of such approaches is to be however incompatible with real-time processing since several frames (7 consecutive frames in [93, 138], the whole 20 frames long sequence in [163]) must be acquired before running the algorithm.

3.4.3 Data sets

The data sets used to validate the proposed methodology introduced in section 3.3 were provided by John Hopkins Applied Physics Laboratory (JHAPL). They were acquired in 2006 at the Dugway Proving Ground in Utah (USA). The recording sensor was a Field-portable Imaging Radiometric Spectrometer Technology (FIRST) [70] long-wave infrared sensor, producing video sequences at a frame-rate of 0.2 Hz, where each frame is a hyperspectral image composed of 128×320 pixels in the spatial domain and 129 spectral channels, spanning 7.81 μm to 11.97 μm in wavelength. Each gas release was recorded by three identical FIRST sensors placed at different locations around the release point, denoted Romeo, Victory and Tango and located 2.15, 2.75 and 2.82 kms away from the release point, respectively, as illustrated by figure 3.7.

The following experiments are conducted on two sequences featuring the explosive re-



(a) (b)

(c) (d)

Figure 3.8: False color RGB composition of (a) the 20$^{\text{th}}$ frame of aa12_Victory and (c) the 22$^{\text{th}}$ frame of aa13_Victory sequences. Corresponding ground truth data (b) and (d), where strongly and weakly concentrated parts are depicted in red and green, respectively.

lease of an acetic acid canister, denoted **aa12_Victory** and **aa13_Victory**, respectively. Despite each sequence initially contained hundreds of hyperspectral frames, only the small portion of the sequence featuring the plume release and diffusion was retained. The resulting aa12_Victory and aa13_Victory sequences are composed of 30 frames each, where only the last 20 frames features the diffusion of the plume (the release therefore occurring at the 11$^{\text{th}}$ frame). Eventually, the radiance for each frame is converted into emissivity using the TES algorithm [29] provided by JHAPL along with the video sequences. Not that, for the following experiments, it is assumed that only the first two frames are known no to contain the gas plume, while the time of release remains unknown.

Ground truth data was created for the two data sets [2] to further evaluate the performance of the proposed tracking algorithm and compare it with two state-of-the-art methods, namely an anomaly detection AMSD method [29] and the robust nonnegative matrix factorization (RNMF) clustering method described in [163]. To generate the ground truth map for a given frame, a principal component analysis (PCA) was performed and the three principal components (PCs) showing the highest contrast between the plume and the background were selected, creating a false color RGB composition of the scene. The PC selection was conducted by visually analyzing the twenty PCs (being most likely to feature some contrast between the gas plume and the background). The identity of the selected PCs varied between the two data sets, and even between two consecutive frames of a single data set, making this selection automated impossible. Two classes were carefully delineated from the RGB composition, the first corresponding to the strongly concentrated section of the plume (typically the central part),

2. Half of this ground truth data was delineated by Delphine Pauwels in the framework of her Master's internship between February and July 2015.

and the second to more diffused components. Figure 3.8 displays examples of ground truth data for both aa12_Victory and aa13_Victory sequences. Due to the inherent subjectivity of the ground truth manual delineation task, it is advocated not to consider the created ground truth as a perfect gold standard, but rather as a support for the quantitative comparison of the performance of several methods.

3.5 Experimental methodology

Gas plume tracking is a challenging task since the gas plume is a non-rigid object with no real boundary and whose shape is evolving quickly and unpredictably. Moreover, the gas plume is an optically thin object whose concentration changes with position and time, as a natural consequence of the diffusion phenomenon, making it more and more difficult to detect. Therefore, an appropriate tuning of the proposed methodology is required in order to efficiently track the diffusing gas plume.

3.5.1 Detection of the release point

As explained in section 3.3.4, the proposed object tracking methodology is triggered once some motion has been detected in the sequence. In the present gas plume tracking application, it means that the moment at which the gas is released in the sequence must be first detected. As it is assumed that the first two frames of the sequences are known not to feature the release of the plume, the change detection procedure described in section 3.3.2.1 is applied from frame #3 on: the difference $\mathcal{I}^{\Delta t} = \mathcal{I}^t - \mathcal{I}^{t-1}$ between the current frame and the previous one is first computed. Each pixel of the frame difference has then its GLRT expression (3.9) computed and thresholded to achieve a probability of detection p_D set to 99%, finally producing the change mask $C_{t-1,t}$. If at least one pixel of the change mask has been detected as changing between $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$, then the tracking procedure is launched, as the size of the plume immediately after the release is unknown and can be considered as arbitrarily small.

To be operable in practice however, this strategy must not produce any false alarm at all, as it would engage the tracking procedure too early otherwise. We noted that the performances of the change detection were slightly enhanced when first transforming the frame difference with a PCA and performing the statistical test (3.9) on the PCs instead of the raw difference. This can be explained by the ability of the PCA transformation to decorrelate the data and hence improve its separability with respect to the highly correlated raw hyperspectral data. Applying a PCA transformation prior to the change detection leads to two modifications in the formulation of the statistical test (3.4). First, the covariance matrix Σ, estimated from the difference of the first two frames of the sequence, is replaced by the diagonal matrix $\Sigma^{\text{PCA}} = \mathbf{W}^{-1}\Sigma\mathbf{W}$, where $\mathbf{W}$ is the eigenvector matrix of Σ. Second, the unknown mean $\boldsymbol{\mu}_i^{\Delta t}$ in the alternate hypothesis $\mathbf{H}_1$ is replaced by $\mathbf{W}^{-1}\boldsymbol{\mu}_i^{\Delta t}$. Up to those modifications, the derivation of the GLRT and the further conclusions drawn in sections 3.3.2.1 and 3.3.4 remain identical.

3.5.2 Motion prediction step

The motion prediction step is conducted as presented in section 3.3.2. First the change mask $C_{t-1,t}$ is computed by thresholding the result of the change detection statistical test between $\mathcal{I}^{t-1}$ and $\mathcal{I}^t$ so it achieves a probability of detection p_D of 95%. Then, the estimate position of the plume $\hat{O}_t$ is obtained from $C_{t-1,t}$ and the previous position of the plume O_{t-1} using equation (3.10). Note that, similarly to what is done for the detection of the release point, the change detection statistical test is performed on the PCA of the frame difference.

3.5.3 Matching step

As described in subsection 3.3.3, the matching step first involves a hierarchical decomposition of the current frame by means of a BPT representation, and then the object detection process using this BPT. As presented in section 1.3.4, the construction of the BPT requires the definition the initial segmentation map, the region model and the merging criterion:

- The *initial segmentation map* is obtained using the hyperspectral watershed algorithm proposed by [188], as its ability to properly delineate the boundaries, even diffused, between the gas plume and the background was demonstrated in [196]. Even if this method is known to produce severe over-segmentation, each hyperspectral frame is initially segmented in around 2000 regions (as compared to the $128 \times 320 = 40960$ potential initial regions if starting the construction of the BPT from the pixel level).
- The *region model* of a region $\mathcal{R}$ is set to the mean spectrum region model (1.13) $\boldsymbol{\mu}_{\mathcal{R}}$, which already proved to perform well to separate the plume from the background [196].
- The *merging criterion* between neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ is defined as the spectral information divergence (1.17) between their region models $\boldsymbol{\mu}_{\mathcal{R}_i}$ and $\boldsymbol{\mu}_{\mathcal{R}_j}$.

The object identification succeeds to the construction of the BPT in order to complete the matching step. It requires the definition of a set of reference features Ω^{O_t} for the tracked object, and the computation of the similarity between the set of features $\Omega^{\mathcal{R}}$ of every region in the BPT and the set of reference features Ω^{O_t} to identify the region that most matches the sought object. The used features and the corresponding distances to compare them are the following:

3.5.3.1 Spectral feature

The proposed *spectral* feature $\omega^{O_t}_{\text{spect}}$ is the mean spectrum $\boldsymbol{\mu}_{O_{t-1}}$ of the plume in the previous frame, being the region model of the region $\mathcal{R} \in H_{t-1}$ that was selected from H_{t-1} to be the object representation O_{t-1}. Between two consecutive frames, the gas plume concentration is expected to limitedly vary. Additionally, the plume general motion is assumed to be slow, meaning that the region contaminated by the plume should overlay similar backgrounds in two consecutive frames and hence having similar mean spectra.

The proposed spectral feature distance is derived from the two-sample Hotelling's T-square statistic, which arises when testing the equality of the mean vectors of two populations [166].

More specifically, let $\hat{\Sigma}_{\mathcal{R}}$ and $\hat{\Sigma}_{O_{t-1}}$ be the respective sample covariance matrices of $\mathcal{R}$ and O_{t-1}. The Hotelling's T-square statistic between $\mathcal{R}$ and O_{t-1} has the following expression:

$$T^2(\mathcal{R}, O_{t-1}) = \frac{|\mathcal{R}||O_{t-1}|}{|\mathcal{R}| + |O_{t-1}|} \left(\boldsymbol{\mu}_{\mathcal{R}} - \boldsymbol{\mu}_{O_{t-1}}\right)^T \Sigma_{\text{pool}}^{-1} \left(\boldsymbol{\mu}_{\mathcal{R}} - \boldsymbol{\mu}_{O_{t-1}}\right) \tag{3.23}$$

with Σ_{pool} being the pooled covariance matrix

$$\Sigma_{\text{pool}} = \frac{(|\mathcal{R}| - 1)\hat{\Sigma}_{\mathcal{R}} + (|O_{t-1}| - 1)\hat{\Sigma}_{O_{t-1}}}{|\mathcal{R}| + |O_{t-1}| - 2}. \tag{3.24}$$

The normalized T^2 statistic

$$F(\mathcal{R}, O_{t-1}) = \frac{|\mathcal{R}| + |O_{t-1}| - N - 1}{N(|\mathcal{R}| + |O_{t-1}| - 2)} T^2(\mathcal{R}, O_{t-1}) \tag{3.25}$$

follows a F-distribution with N numerator degrees of freedom and $|\mathcal{R}| + |O_{t-1}| - N - 1$ denominator degrees of freedom. The final spectral feature distance d_{spect} is finally obtained by normalizing the F-statistic between $\mathcal{R}$ and O_{t-1} with the F-statistic between $\hat{O}_t$ and O_{t-1}

$$d_{\text{spect}}\left(\omega_{\text{spect}}^{\mathcal{R}}, \omega_{\text{spect}}^{O_t}\right) = \frac{F(\mathcal{R}, O_{t-1})}{F(\hat{O}_t, O_{t-1})} \tag{3.26}$$

as it is desirable for a good candidate region to be closer from the reference spectrum $\boldsymbol{\mu}_{O_{t-1}}$ than the estimate region $\hat{O}_t$. Distance values for those good candidate regions thus range between 0 and 1. Note that the feature distance can be computed only for regions $\mathcal{R}$ whose size $|\mathcal{R}|$ is greater than the number of spectral bands N to correctly estimate the sample covariance matrix $\hat{\Sigma}_{\mathcal{R}}$.

3.5.3.2 Spatial feature

The proposed *spatial* feature $\omega_{\text{spat}}^{O_t}$ is the binary position and shape of $\hat{O}_t$. It is indeed assumed that, even if perfectible, the output $\hat{O}_t$ of the motion prediction is a good initial guess for the spatial position and shape of the plume in $\mathcal{I}^t$.

Consequently, the proposed spatial feature distance evaluates how similar from $\hat{O}_t$ is any candidate region $\mathcal{R}$:

$$d_{\text{spat}}\left(\omega_{\text{spat}}^{\mathcal{R}}, \omega_{\text{spat}}^{O_t}\right) = \frac{|\mathcal{R}\Delta\hat{O}_t|}{|\mathcal{R}|} \tag{3.27}$$

where $|\mathcal{R}\Delta\hat{O}_t|$ is the number of pixels in the symmetric difference between $\mathcal{R}$ and $\hat{O}_t$, *i.e.*, pixels either in $\mathcal{R}$ or in $\hat{O}_t$, but not in both. Therefore, the spatial distance (3.27) corresponds to the percentage of pixels of $\mathcal{R}$ that mismatch $\hat{O}_t$. Thus, the closer from $\hat{O}_t$ in terms of position and shape is $\mathcal{R}$, the smaller this percentage of error.

3.5.3.3 Temporal feature

The proposed *temporal* feature $\omega_{\text{temp}}^{O_t}$ is defined as a confidence area where the tracked plume is expected to be found with certainty. This confidence area is derived from the estimate position $\hat{O}_t$ by a morphological dilation with a structuring element SE, $\delta_{\text{SE}}(\hat{O}_t)$, and the percentage of inclusion of every candidate region $\mathcal{R}$ in the confidence area is evaluated:

$$\mathcal{R}_{\%\hat{O}_t} = \frac{|\mathcal{R} \cap \delta_{\text{SE}}(\hat{O}_t)|}{|\mathcal{R}|} \tag{3.28}$$

The proposed temporal feature distance is a hard thresholding of this percentage of inclusion:

$$d_{\text{temp}}\left(\omega_{\text{temp}}^{\mathcal{R}}, \omega_{\text{temp}}^{O_t}\right) = \begin{cases} 0 & \text{if } \mathcal{R}_{\%\hat{O}_t} \geq \tau \\ +\infty & \text{otherwise} \end{cases} \tag{3.29}$$

This distance allows to consider only regions in the BPT that have at least $\tau\%$ of their pixels in the confidence area as possible candidates, dismissing all other regions. It is possible to be more or less selective by varying the structuring element SE and the threshold τ. In practice, we used a 9×9 square structuring element, and τ was set to 80%.

3.5.3.4 Final matching distance

For each region $\mathcal{R}$ of the BPT H_t, the global distance to the set of reference features Ω^{O_t} is finally obtained by adding the three feature distances:

$$d\left(\mathcal{R}, O_t\right) = d_{\text{spect}}\left(\omega_{\text{spect}}^{\mathcal{R}}, \omega_{\text{spect}}^{O_t}\right) + d_{\text{spat}}\left(\omega_{\text{spat}}^{\mathcal{R}}, \omega_{\text{spat}}^{O_t}\right) + d_{\text{temp}}\left(\omega_{\text{temp}}^{\mathcal{R}}, \omega_{\text{temp}}^{O_t}\right) \tag{3.30}$$

where d_{spect}, d_{spat} and d_{temp} are defined following equations (3.26), (3.27) and (3.29), respectively. Note that this is equivalent to only considering the spectral and spatial feature distances for regions whose temporal feature distance is equal to 0. The optional weights of equation (3.11), which trade off the influence of a distance with respect to the other, are all set to 1 because the spectral and spatial feature distances have a similar range of values for good candidate regions. The optimal region $\mathcal{R}^\star$ of the BPT H_t is finally the one achieving the smallest global distance to the set of reference features, $\mathcal{R}^\star = \text{argmin}_{\mathcal{R} \in H_t} d\left(\mathcal{R}, O_t\right)$ and is retrieved with an exhaustive search.

3.5.4 The adaptive matched subspace detector

In order to validate the proposed methodology, we compare it with the state-of-the-art AMSD algorithm that was notably investigated in [29, 30, 147] for the detection of chemical gas plume in hyperspectral video sequences. Each frame of the sequence is processed independently of the others. The AMSD involves a two-hypotheses testing, performed on the raw hyperspectral

frame, considering target pixels as anomalies with respect to a structured background model. More specifically, the two competing hypotheses are

$$\begin{aligned} \mathbf{H}_0 : \quad & \mathbf{x}_i^t = \mathbf{B}^t \boldsymbol{\beta}_i^t + \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \text{ (target absent)} \\ \mathbf{H}_1 : \quad & \mathbf{x}_i^t = \mathbf{t}\alpha_i^t + \mathbf{B}^t \boldsymbol{\beta}_i^t + \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \text{ (target present)} \end{aligned} \tag{3.31}$$

where $\mathbf{B}^t$ is a (possibly varying with time) $N \times Q$ matrix composed of Q background emissivity signatures, $\boldsymbol{\beta}_i^t$ is a $Q \times 1$ vector containing the respective weights of the background emissivity signatures for pixel $\mathbf{x}_i^t$, $\mathbf{t}$ is the $N \times 1$ *a priori* known target vector, α_i^t is the relative weight of the target emissivity contained in the current pixel, and the additive noise is assumed to be Gaussian with zero mean and $\sigma^2\mathbf{I}$ covariance. The use of the GLRT approach leads to the following statistical test:

$$\Lambda_{\text{AMSD}}(\mathbf{x}_i^t) = \frac{(\mathbf{x}_i^t)^T \left(\mathbf{P}_{B^t}^{\perp} - \mathbf{P}_Z^{\perp} \right) \mathbf{x}_i^t}{(\mathbf{x}_i^t)^T \mathbf{P}_Z^{\perp} \mathbf{x}_i^t} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{AMSD}} \tag{3.32}$$

with $\mathbf{Z} = [\mathbf{B}^t \ \mathbf{t}]$ being the concatenation of the background emissivity and target emissivity matrices and $\mathbf{P}_A^{\perp}$ is the projection matrix of $\mathbf{A}$ defined by $\mathbf{P}_A^{\perp} = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, with $\mathbf{A} = \mathbf{B}^t$ and $\mathbf{A} = \mathbf{Z}$, respectively. It is known [131] that

$$(N - 1 - Q)\Lambda_{\text{AMSD}}(\mathbf{x}_i^t) \sim F_{1,N-1-Q}(\text{SINR}_o) \tag{3.33}$$

where $F_{1,N_\lambda-1-Q}(\text{SINR}_o)$ is the non-central F distribution with 1 numerator degree of freedom, $N - 1 - Q$ denominator degrees of freedom, and non-centrality parameter SINR_o defined as

$$\text{SINR}_o = \frac{\left\| \mathbf{P}_{B^t}^{\perp} \mathbf{t} \alpha_i^t \right\|^2}{\sigma^2} \tag{3.34}$$

Under $\mathbf{H}_0$, $\text{SINR}_o = 0$ and the AMSD statistic depends only on the known parameters N and Q and enjoys a constant false alarm rate property. The Q background emissivity signatures, needed to operate the AMSD, are estimated as the first Q principal components of a set of reference background pixels [29,233]. This set of reference is firstly defined as the first frame of the sequence, on which a PCA is performed and the first Q PCs are extracted. For each new incoming frame, the AMSD statistic is run on each pixel, and all pixels whose test statistic is below a predefined threshold δ (the most likely to follow $\mathbf{H}_0$) are added to the set of reference background pixels. This updated set is used to generate the Q background emissivities for the next frame. We used the AMSD implementation and the reference target emissivity spectrum $\mathbf{t}$ provided by the JHAPL along with the data sets. $Q = 4$, $\delta = 0.25$ and a probability of false alarm $p_{FA} = 5\%$ to compute γ_{AMSD} under $\mathbf{H}_0$ were used in the implementation, as advocated in [29].

3.5.5 The robust nonnegative matrix factorization clustering

The second state-of-the-art method we compare the proposed methodology with is the robust non-negative matrix factorization (RNMF) clustering method proposed in [163]. Representing the hyperspectral frame $\mathcal{I}^t$ by a $N_x N_y \times N$ matrix $\mathbf{X}$, the RNMF clustering method

aims at decomposing $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{L} + \mathbf{S} \tag{3.35}$$

where $\mathbf{L}$ is a low rank matrix accounting for the data, and $\mathbf{S}$ is a sparse matrix representing the noise corrupting the data. Further, the matrix $\mathbf{L}$, corresponding to the data points, is decomposed as

$$\mathbf{L} = \mathbf{YZ}, \qquad \mathbf{Y}, \mathbf{Z} \geq 0 \tag{3.36}$$

where $\mathbf{Z}$ is the collection of all cluster centroids, and $\mathbf{Z}$ is the cluster indicator matrix. The final RNMF clustering method can finally be written as the solution of the following optimization problem:

$$\min_{\mathbf{Y},\mathbf{Z}\geq 0,\mathbf{X}=\mathbf{L}+\mathbf{S}} \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \frac{\rho}{2}\|\mathbf{L} - \mathbf{YZ}\|_F \tag{3.37}$$

with $\|\mathbf{L}\|_*$ being the nuclear norm of $\mathbf{L}$ (*i.e.*, the sum of all singular values of $\mathbf{L}$, enforcing $\mathbf{L}$ to be low-rank), $\|\mathbf{S}\|_1$ being the L_1 norm of the noise matrix $\mathbf{S}$ (therefore favoring sparsity), and $\|\mathbf{L} - \mathbf{YZ}\|_F$ being the Frobenius norm of the error between the low-rank data matrix $\mathbf{L}$ and its approximation $\mathbf{YZ}$, where both $\mathbf{Y}$ and $\mathbf{Z}$ are expected to have non-negative entry values. The optimization problem (3.37) is solved by the alternating direction method of multipliers [118].

The RNMF results presented in the following section 3.6 are reproduced from [163], where the clustering method (3.35) was applied both on aa12_Victory and aa13_Victory sequences. To account for the temporal evolution of the gas plume along the sequence, 20 frames (2 frames prior to the release of the plume and the following 18 frames featuring the diffusion for aa12_Victory, 1 frame prior to the release and the following 19 frames with the plume for aa13_Victory) were first stacked together prior to the solving of the optimization problem (3.37), which does not allow for this strategy to be operated in practice in real-time scenarios. Note also that this method requires the number of final clusters as an input. This number was set to 4 (being the sky, the foreground, the mountain and the plume).

3.6 Results

3.6.1 Assessing the tracking quality

Assessing the performances of a tracking algorithm is a well-known challenge in computer vision. Several studies have addressed the problem when ground truth data is available. In [145], a metric is introduced to compare the trajectory of the tracked object with a reference trajectory accounting for ground truth. In [25, 231], frame-based surveillance metrics relying on the number of true and false positives are developed. These metrics are notably used to evaluate the consistency of the tracker across the whole sequence (where a true positive is claimed when the object of interest is present in a given frame and correctly detected by the tracker). Object-based performance metrics, such as spatial overlap between ground truth object and tracked object and Euclidean distance between their respective centroids, are considered in [16]. In particular, we propose to use this notion of overlapping between ground truth and corresponding object in order to derive three metrics reflecting the performance and

accuracy of the tracking. As can be seen in figure 3.8, the ground truth map for each frame is composed of three different regions:

- Regions where the plume is strongly concentrated (in red in figures 3.8b and 3.8d), denoted $\mathrm{GT}_{\mathrm{strong}}$ in the following.
- Regions where the plume is weakly concentrated (in green in figures 3.8b and 3.8d), denoted $\mathrm{GT}_{\mathrm{weak}}$.
- All remaining regions of the image, displayed in blue in figures 3.8b and 3.8d, not containing any gas and denoted $\mathrm{GT}_{\emptyset}$.

We define the percentage of strong detections N_{sd} and of weak detections N_{wd} as the percentage of strongly and weakly concentrated ground truth plume areas included in O_t, respectively:

$$N_{sd} = \frac{|O_t \cap \mathrm{GT}_{\mathrm{strong}}|}{|\mathrm{GT}_{\mathrm{strong}}|} \tag{3.38}$$

and

$$N_{wd} = \frac{|O_t \cap \mathrm{GT}_{\mathrm{weak}}|}{|\mathrm{GT}_{\mathrm{weak}}|}. \tag{3.39}$$

Similarly, the percentage of false alarms N_{fa} is defined as the percentage of $\mathrm{GT}_{\emptyset}$ area that is wrongly comprised in O_t,

$$N_{fa} = \frac{|O_t \cap \mathrm{GT}_{\emptyset}|}{|\mathrm{GT}_{\emptyset}|}. \tag{3.40}$$

High values of N_{sd} and N_{wd} (theoretically, 1) along with a low value of N_{fa} (theoretically, 0) indicate a good detection of the plume for a given frame. The temporal performance of the tracking can be assessed by evaluating the consistency of N_{sd} and N_{wd} to remain high and of N_{fa} to stay low across the whole sequence.

3.6.2 Results

Quantitative results for aa12_Victory and aa13_Victory sequences are presented by figures 3.9 and 3.11, respectively. Each figure is composed of three plots representing the evolution of the percentages of strong detections, weak detections and false alarms across the sequence, where the x-axis correspond to the frame number and the y-axis is the percentage. Each plot is composed of a red solid line and two dashed black and blue line curves, where the former corresponds to the proposed object tracking and the latter are the result of state-of-the-art AMSD and RNMF methods, respectively. In addition to the quantitative results, qualitative results for the aa12_Victory and aa13_Victory sequences are displayed by figures 3.10 and 3.12, respectively. Each figure is composed of ten rows and four columns. Each row presents, from left to right, the RGB representation of a hyperspectral frame of the sequence and the binary mask of the detected plume for the proposed BPT-based and state-of-the-art AMSD and RNMF methods, respectively. Each column corresponds to a particular frame of the sequence. While it is impossible to show all frames by lack of room, only frames #11, #12, #14, #16, #18, #20, #22, #24, #26 and #28 are represented.

We recall that the release of the plume occurs in frame #11 in both sequences. For the RNMF method, only 20 frames among the 30 that constitute the sequence are available.

3.6.2.1 The aa12_Victory sequence

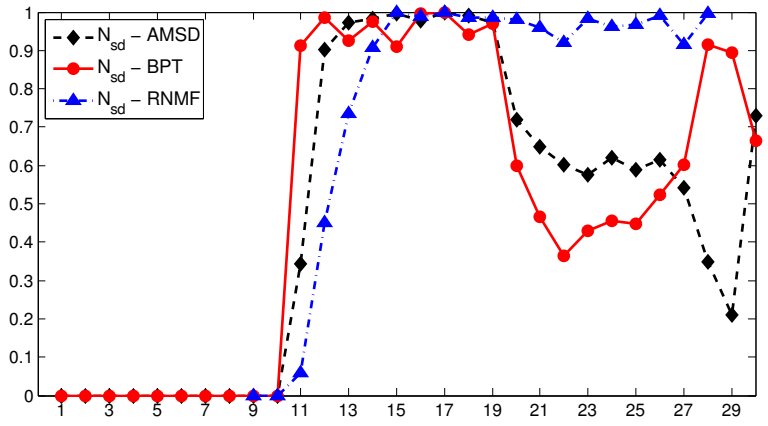
About the strong detection plot The behavior of the evolution of the N_{sd} values the aa12_Victory sequence is show by figure 3.9a, and several observations arise from its analysis.

First, all AMSD, BPT and RNMF methods are able to correctly detect the frame where the gas is released, as they all score a N_{sd} value greater than 0 for the 11^{th} frame. Nevertheless, the percentage of strong detections for the RNMF method is under 10%, and one can see in figure 3.10 that only a small portion of the released plume is detected.

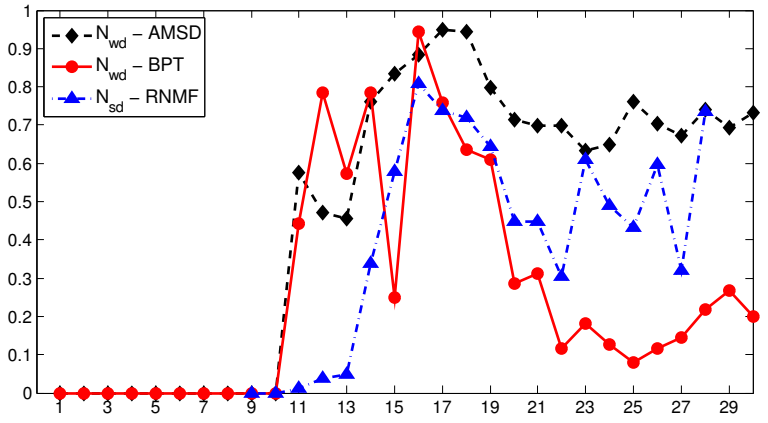
Then, the evolution of the N_{sd} curves for the AMSD and proposed BPT-based methods feature a similar trend: remaining over 90% of strong detections until frame #19, they both suffer a drop from frame #20 on, before rising again at the end of the sequence. This decrease can be explained by the fact that at this point of the sequence, the plume equally overlays the foreground and the sky, as shown by figures 3.8a and 3.10, and both methods suffer from this split. For the proposed BPT-based method in particular, both halves of the plume (the one overlaying the foreground and the other covering the sky) are supported by nodes in different branches in the hierarchical decomposition of the frames. As the object detection process was implemented to extract only one node from the hierarchical decomposition to represent the tracked object during the matching step, half of the plume is lost. This notably explains why the percentage of strong detection N_{sd}-BPT is divided by 2 between frames #19 and #22. When evaluating the spectral feature during the object identification process, the reference spectrum before the split corresponds to the plume mainly overlaying the foreground. Therefore, the mean spectrum of the half of the plume overlaying the foreground directly after the split is closer to the reference spectrum than the one of the half overlaying the sky. The bottom half of the plume is subsequently correctly tracked by the proposed method, as it can be observed on figure 3.10. The percentage of strongly concentrated plume N_{sd} is increasing again for the proposed BPT-based method from frame #25 on because the top half of the plume is gradually disappearing from the frame of view, leaving only residuals that are classified as weakly concentrated in the ground truth map.

The behavior of N_{sd}-RNMF is slight different from the other two methods. While it remains consistently over 90% between frame #15 and frame #28 (the last available one for this method), it gradually increases for the first 4 frames of the sequence, as if the clustering method was unable to efficiently differentiate between the plume signature and the background for these initial frames of the sequence. A possible, yet rather surprising, explanation for this observation is that the plume is "too much" concentrated in those frames. As the RNMF clustering method is operating on the 20 frames stacked together, it may consider that those highly concentrated plume pixels do not belong to the same cluster as the more diffused plume pixels, and mis-classifies them as foreground.

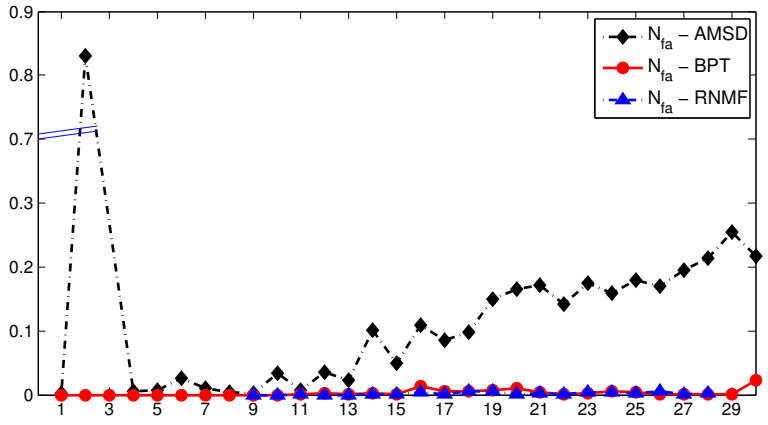
About the weak detection plot The evolution of the weak detection percentages N_{wd} for the aa12_Victory sequence is displayed by figure 3.9b. The ground truth data labeled as weakly concentrated corresponds to areas where the plume is diffused and thin, and is



(a)



(b)



(c)

Figure 3.9: Percentage of (a) strong detection, (b) weak detection and (c) false alarms for the aa12_Victory data set. Dashed black and blue lines correspond to state-of-the-art AMSD and RNMF, respectively, while plain red line corresponds to the proposed BPT-based method. For the false alarm plots (c), the y-axis has been broken for an easier visualization and comparison of the two method performances.

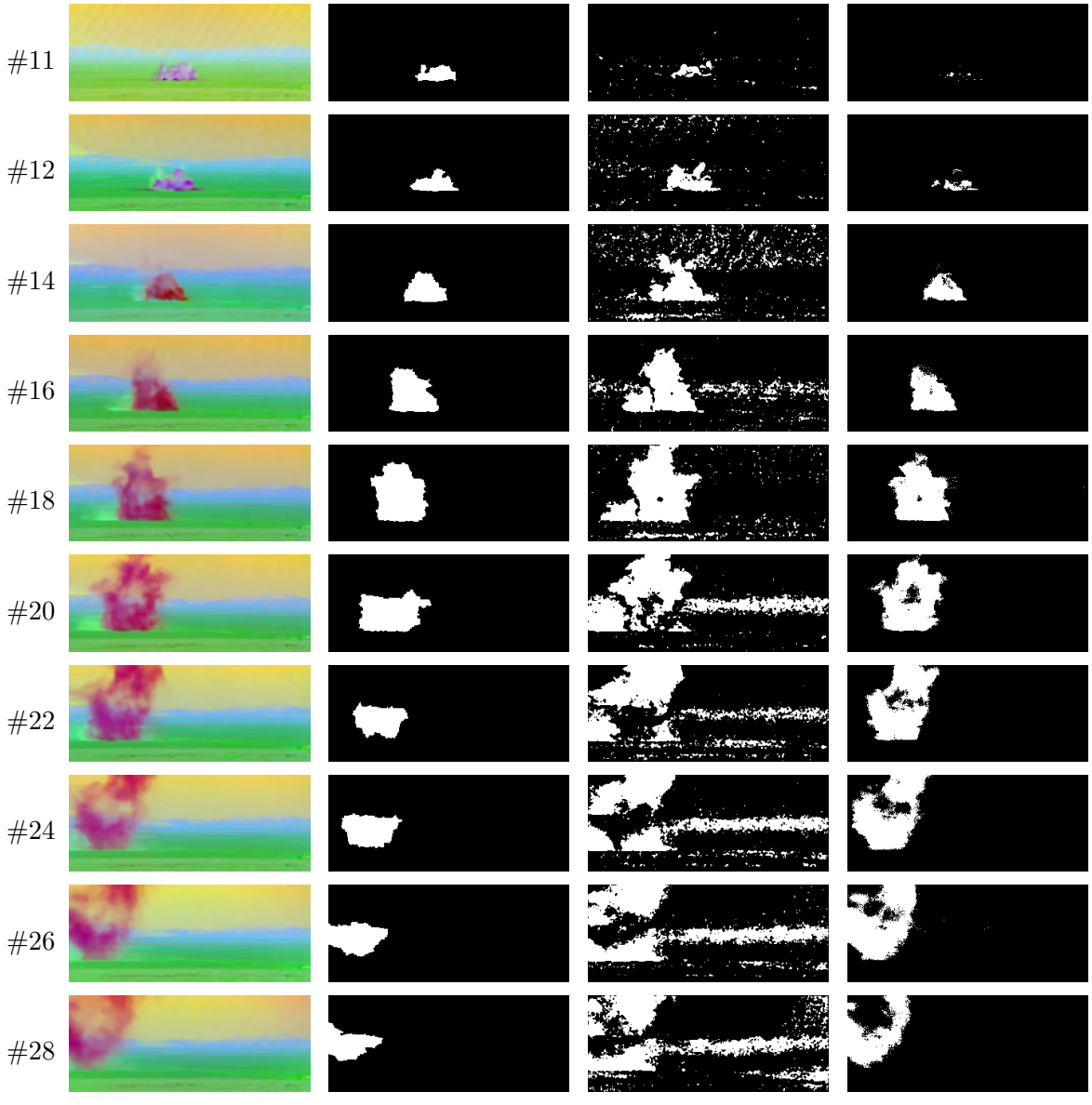


Figure 3.10: Visual results of the tracking for 10 frames of aa12_Victory sequence. Column (a) shows the RGB representation of the hyperspectral frame, and columns (b), (c) and (d) show the binary mask of the detected plume for the proposed BPT-based, the AMSD and the RNMF methods, respectively.

very likely to be mistaken with the background. It is the reason why the weak detection percentages are lower than the strong detection percentages for all three methods and the majority of the frames.

However, there are some common patterns with the strong detection plot. For the proposed BPT-based method, the N_{wd} value remains relatively high for the first half of the frames (except for frame #15) and significantly drops for the second half. This is in accordance with the explanation that the plume is splitting over the sky and the foreground, and only the bottom half is tracked. The top part, overlaying the sky, quickly diffuses and is labeled as weakly concentrated in the ground truth, therefore missed by the tracker. As for the strong detections, the weak detection values N_{wd} for the state-of-the-art RNMF clustering remains low for the first 4 frames featuring the release, confirming that the method does not assign those pixels (even if moderately concentrated) to the plume cluster.

About the false alarm plot The percentage of false alarms N_{fa} for the aa12_Victory data set is presented by figure 3.9c. The y-axis has been cut for a better visualization. The false alarms plots shall be analyzed in two time intervals: prior to and after the release of the plume.

For the first case, one can see by observing the N_{fa}-BPT plot, that the percentage of false alarms remain equal to 0 prior to the release. This implies that the the proposed method does not generate any false alarm before the appearance of the plume, meaning that the tracking algorithm is triggered at the right time. It demonstrates the robustness of the change detection hypothesis testing, performed on the PCA transformation of the frame difference, even with a high probability of detection threshold (being set to $p_D = 99\%$ prior to the detection of the release). In comparison, the AMSD method consistently produces false alarms before the release of the plume. While the N_{fa} remains low for most frames prior to the release, it achieves a peak over 80% for frame #2. This could be problematic in a scenario where further processings relying on a precise detection of the plume release are needed.

The RNMF and proposed BPT-based method only generate a tiny amount of false alarms (no more than 2% of all background pixels) for all frames. While not necessarily implying that all plume pixels are correctly detected, this observation signifies that background pixels are however not confused with the plume. For the proposed method in particular, it indicates that the BPT is able to properly separate the plume from the background during the construction of the tree, suggesting that the mean spectrum, even if relatively standard, is an appropriate choice for the region model. For the AMSD however, the amount of false alarms is globally increasing, up to over 20% for the last frames of the sequence, meaning that the method is over-estimating the plume when this latter becomes diffused. Looking in particular at figure 3.10, one can see that the AMSD detect as plume both the small cloud of dust which has been triggered by the explosive release (appearing as bright green on the left hand side of the red plume) and part of the mountain in the background.

3.6.2.2 The aa13_Victory sequence

About the strong detection plots The percentage of strong detections N_{sd} over the aa13_Victory sequence for the three investigated methods is displayed by figure 3.11a.

The first observation that can be made is that the two state-of-the-art AMSD and RNMF methods miss the release of the plume occurring at the 11^{th} frame, while the proposed BPT-based method scores almost 90% of strong detections for this frame. This demonstrates again the robustness of the implemented change detection approach to perform the tracking initialization, and the subsequent object detection procedure, as the proposed method is able to lock on the plume directly as it appears in the sequence.

Contrarily to the aa12_Victory sequence, the behavior of N_{sd}-BPT plot this time resembles this of N_{sd}-RNMF. As a matter of fact, they both remain over 70% of strong detections until frame #25 before significantly dropping for the last frames of the sequence. More particularly, the percentage of strong detections for the proposed BPT-based method even drops to 0% for the last two frames #29 and #30, meaning that the track has been completely lost. The main reason is that the gas has become so diffused as this point of the sequence that the change detection test no longer detects change between two consecutive frames, resulting in an empty change mask $C_{t-1,t}$. Consequently, the output $\hat{O}_t$ of equation (3.10) when $C_{t-1,t}$ is empty is equal to the previous object position O_{t-1}. In that case, the motion prediction step becomes trapped in one part of the image and the track is lost. Figure 3.12 shows in particular that the tracker is being locked on the background mountain. A possible solution to overcome this issue would be to regularly reset the motion predictor by estimating the change between the current frame and a frame prior tot the release instead of using two consecutive frames and a more limited motion.

The evolution of the N_{sd}-AMSD plot is opposite to the two other curves, dropping at the middle of the sequence before increasing again, outperforming the RNMF and BPT-based methods for the last 5 frames of the sequence, scoring consistently over 95% of strong detections.

About the weak detection plots The evolution of the weak detection percentages N_{wd} for the aa13_Victory data set is shown by figure 3.11b. For the 11^{th} frame, the percentage of weak detection for the RNMF method is also equal to 0%, confirming that the method misses the release of the plume. It is however slightly above 0% for the AMSD, but nevertheless below its number of false alarms N_{fa} for this frame, suggesting that the method is not able either to properly detect the release of the plume. This is confirmed by the analysis of figure 3.12.

For all three methods, the evolution of the weak detection percentages is more chaotic than the strong detections, as the curves oscillates a lot and do not show much consistency from one frame to the other. One can however remark a peak centered around the 25^{th} and 26^{the} frames for the proposed BPT-based methods. As a matter of fact, the method has just lost the track of the object and is settling in a particular region of the image instead. In frame #26, this region coincide with the wake of the plume, explaining the spike in the curve.

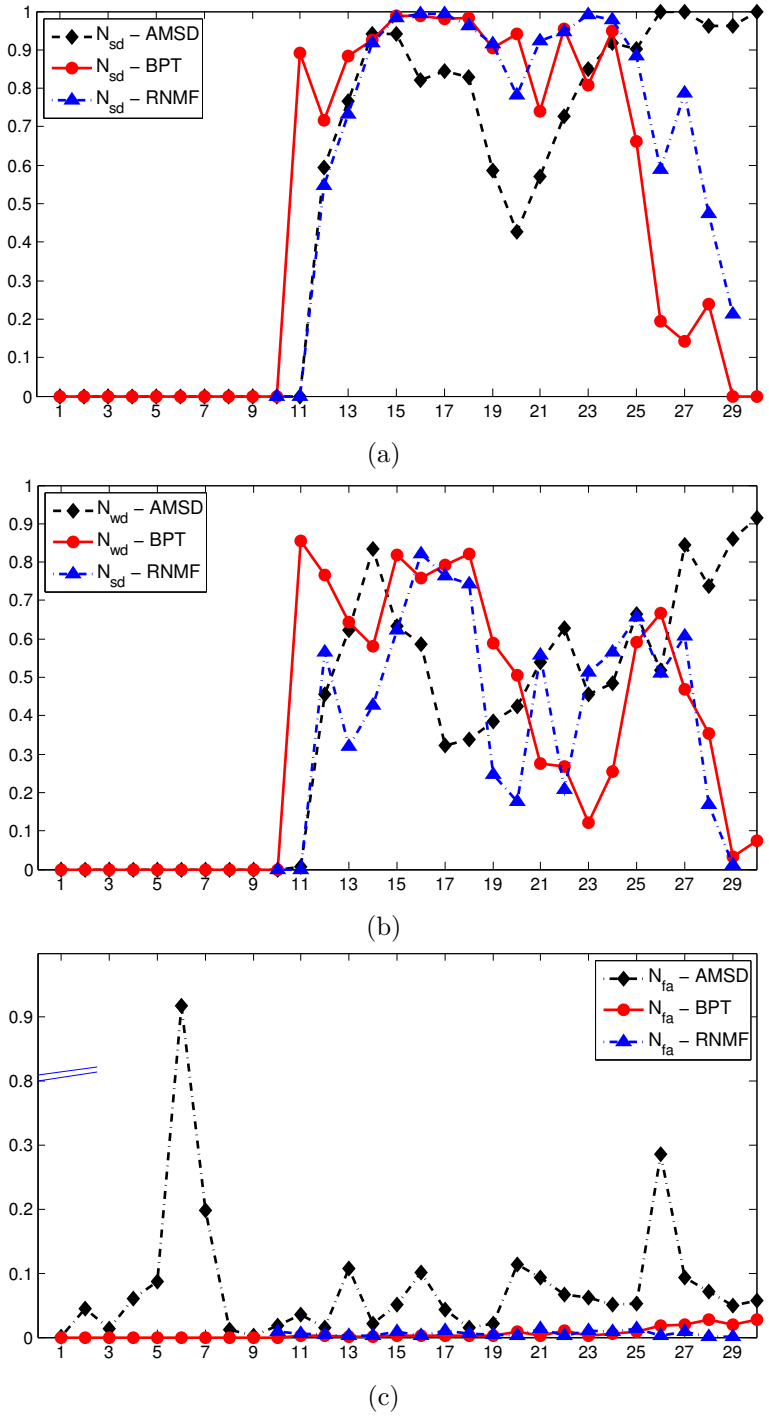


Figure 3.11: Percentage of (a) strong detection, (b) weak detection and (c) false alarms for the aa13_Victory data set. Dashed black and blue lines correspond to state-of-the-art AMSD and RNMF, respectively, while plain red line corresponds to the proposed BPT-based method. For the false alarm plots (c), the y-axis has been broken for an easier visualization and comparison of the two method performances.

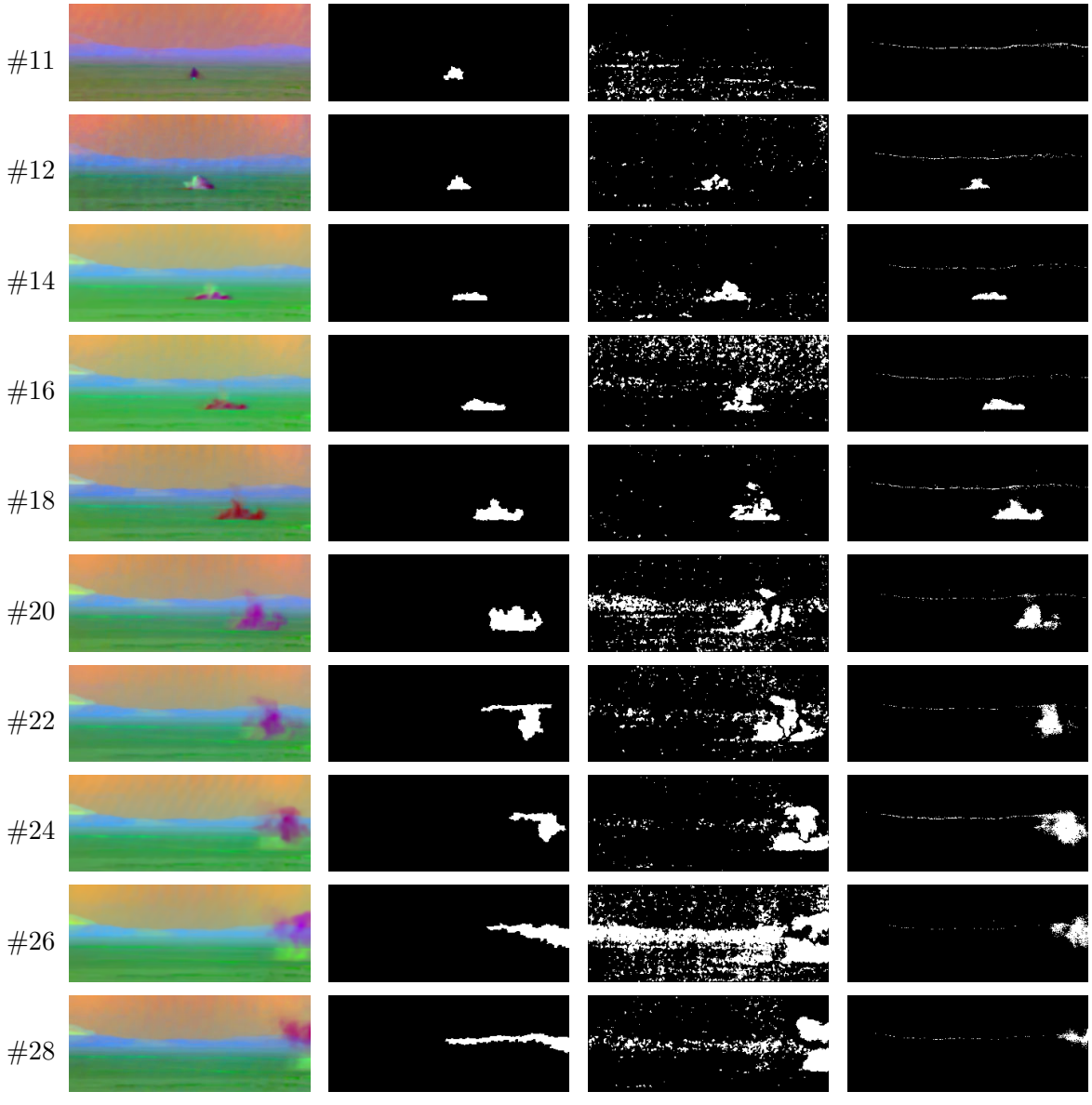


Figure 3.12: Visual results of the tracking for 10 frames of aa13_Victory sequence. Column (a) shows the RGB representation of the hyperspectral frame, and columns (b), (c) and (d) show the binary mask of the detected plume for the proposed BPT-based, the AMSD and the RNMF methods, respectively.

About the false alarm plots Figure 3.11c exhibits the percentage of false alarms of the AMSD, RNMF and BPT-based methods for the aa13_Victory sequence. The y-axis has been cut for a better visualization. Similarly to the aa12_Victory data set, the percentage of false alarms should be analyzed in two parts, being prior to and after the release of the plume in the sequence.

For the first 10 frames, the observations are similar to those made for the aa12_Victory sequence. While the AMSD generates false alarms for all frames prior to the release (with a maximum over 90% for the 6^{th} frame, meaning that almost all pixels of this frame are detected as anomalous by the AMSD), the proposed BPT-based method remains at 0%, validating the fact that the tracking algorithm is not triggered prior to the release.

For all remaining 20 frames, the RNMF and the proposed method do not generate a lot of false alarm, confirming that both methods are able not to confuse the background pixel with the plume. The loss of track for the BPT-based method is illustrated by the increase in percentage of false alarms from frame #25 on, remaning nevertheless moderate (no more than 5%). However, the AMSD produces again a high number of false alarms, consistently over-estimating the plume. Looking at figure 3.12 reveals that, similarly to the aa12_Victory data set, the AMSD tends to label as plume the bright green cloud of dust as well as part of the background moutain. A possible solution to this issue would be to decrease the probability of false alarm p_{FA} (here set to 5%) under which the AMSD statistic is operated, at the expense of the probability of detection, hence lower N_{sd} and N_{wd} values.

3.7 Conclusion

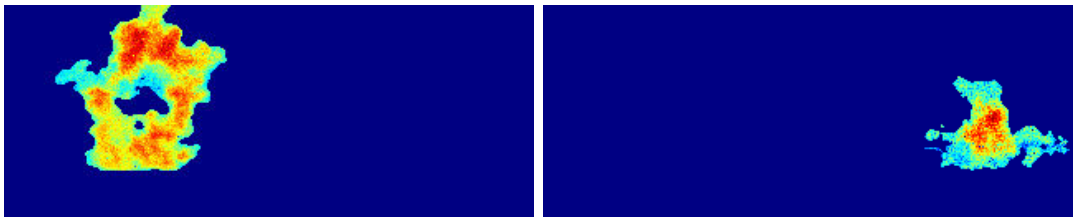
In this chapter, we have investigated the temporal multimodality, in the form of sequences of images collected at close acquisition times. Such sequence of gray-scale and color images, more traditionally termed video sequences, have been the concern of extensive consideration in the field of computer vision. Alternately, hyperspectral video sequences acquired at near real time frame rates have received much less attention. However, such sequences are of interest for several practical real life scenarios, and the need for efficient processing algorithms is growing. In particular, we focused in this chapter on the object tracking application, which is the process of following the motion of objects of interest as they evolve with time along the sequence. While the literature features plenty of object tracking algorithms for traditional video sequences, those methods poorly adapt to the high dimensionality of hyperspectral data, hence the need for adapted algorithms.

Therefore, we proposed a novel methodology to perform hyperspectral object tracking, based on the hierarchical analysis of hyperspectral video sequences, the only prerequisite for the sequence being that only the object of interest is in motion over a fixed background. Like classical object tracking algorithms for color video sequences, the proposed method was decomposed in a motion prediction step and a matching step, performed sequentially. The motion prediction, made of two inner stages, first involved the derivation of the change mask $C_{t-1,t}$ between consecutive hyperspectral frames $\mathcal{I}^{t-1}$ and $\mathcal{I}^{t}$. Writing each pixel signature

as the linear combination of the moving object and the fixed background, this led us to formulate the change detection process as a two-hypotheses statistical test whose solving was conducted with the derivation of a Generalized Likelihood Ratio Test. The resulting change mask was then combined with the known position of the object in the previous frame to derive an estimate of its new position in the current frame. The matching step was defined as an object detection process. The use of a hierarchical decomposition to that purpose allowed to drastically reduce the object search space by representing a hyperspectral frame as a limited set of hierarchically organized candidate regions. The matching involved the definition of a set of reference features for the sought object and the evaluation of every candidate region features against those reference ones, in order to retrieve among the hierarchy, defined as a BPT representation in this chapter, the candidate region matching the tracked object the best, this region subsequently being defined as the object instance in the current frame.

The proposed method was applied to the tracking of chemical gas plumes in two different LWIR hyperspectral video sequences. This challenging application is up to now addressed in the literature either through the use of anomaly detectors or with clustering-based methods. In order to compare the performances of our proposed methodology with two state-of-the-art methods being the Adaptive Matched Subspace Detector (belonging to the class of anomaly detector approaches) and the Robust Non-negative Matrix Factorization-based clustering (therefore falling in the scope of clustering-based techniques), ground truth data was manually delineated for all frames of the two sequences featuring the diffusion of the plume, and corresponding performance metrics were introduced. The conducted performance evaluation showed that the proposed BPT-method was outperforming the AMSD in terms of accurate detection of the gas plume and in the ability not to produce any false alarm detections prior to the release, and was performing equally with the RNMF clustering method for the whole aa13_Victory and the first half of the aa12_Victory sequences (then suffering in the latter case from the assumption that the tracked object shall be represented by a single region in the hierarchical decomposition). Moreover, it is noteworthy that the AMSD method requires the knowledge of the target gas spectrum in order to be operated, and this information may not be available in practical surveillance applications. Similarly, it is worth recalling that, even if producing the best results, the RNMF clustering method needs to stack several consecutive frames at once in order to take into account some temporal information. More particularly, 20 frames were all stacked together for both sequences prior to the clustering, this strategy being incompatible with a real time operation.

Several aspects of the proposed methodology deserve some additional attention in the future in order to improve its robustness with respect to practical hyperspectral object tracking scenarios. From a methodological point of view, the assumption that only the object to track is in motion in the sequence should be relaxed in order to extend the proposed approach to the simultaneous tracking of several objects. In addition, those object should also be allowed to split in several nodes in the BPT representation without impeding the tracking. Related to the gas plume tracking application, all state-of-the-art methods only provide information related to the position and shape of the gas plume, *i.e.*, all results can be visualized as a binary detection map where "true" pixel corresponds to the plume. A further step could be to also include some knowledge related to the concentration of those plume pixels. A



(a) (b)

Figure 3.13: Examples of abundance information in addition to the plume detection for (a) the 20^{th} frame of aa12_Victory sequence and (b) the 22^{th} frame of aa13_Victory sequence. A red value signifies a high concentration of the plume.

first attempt to incorporate a spectral unmixing step within the tracking procedure has been recently conducted [3]. The preliminary obtained results, very promising as it can be seen on figure 3.13, are currently being prepared for submission, and encourage us to engage additional efforts in this way.

3. This has been investigated by Delphine Pauwels during her Master's internship conducted within the GIPSA-lab between February and July 2015.

Chapter 4

Sensorial multimodality

Contents

In this chapter, we investigate the sensorial multimodality, that is, when several images of the same scene are acquired with different sensor types. Hence, each individual modality features some particular aspects of the imaged scene resulting from the intrinsic specificities of its associated sensor. The broad variety of imaging sensors induces a large number of possible sensorial multimodal images, and makes the design of generic algorithms to process them very challenging. On the other hand, numerous typical image processing applications would surely benefit from the conception of such sensorial multimodal processing, and image segmentation is one of them As a matter of fact, image segmentation, as it aims as dividing up the image into regions that "make sense", could make the most of both the inherent redundancy and complementarity information that is contained in the multimodal image in order to produce more accurate segmentation maps. However, the generic integration of this multimodal information for segmentation purposes appears as a real challenge in terms of practical information fusion as well as on the adaptability with respect to all possible multimodal configurations. Embedding the use of this multimodal information within a hierarchical analysis framework raises on the other hand the question on the combination of

several hierachies. Recently however, the concept of braids of partitions has been proposed as a theoretical tool and possible solution to this fusion of hierarchies issue, but has not been implemented in practice yet. In this chapter, we develop a fully novel methodology to achieve the segmentation of multimodal images, based on this notion of braids of partitions and formulated in an energetic framework. The remainder of this chapter is organized as follows: section 4.1 introduces the sensorial multimodality and the challenges that have to be faced by multimodal segmentation processings. In section 4.2, we extend the properties related to hierarchical energy minimization introduced in chapter 2 and reformulate them with a lattice terminology, following the seminal work of Kiran [101, 103]. Section 4.3 defines braids of partitions and their subsequent energy minimization prcedure. Section 4.4 features the proposed methodology, namely guidelines to construct a braid from multiple hierarchies, and how to derive an appropriate multimodal segmentation from it. The proposed procedure is tested and validated in section 4.5 on two multimodal data sets with different characteristics. Section 4.6 concludes the present chapter.

A draft of the materials presented in this chapter have been published at the International Conference on Mathematical Morphology (ISMM) 2015 [197], and an extended version, on which this chapter is based on, is currently under review in an international journal [198]. This work has been partially funded through the ERC CHESS project, ERC-12-AdG-320684-CHESS.

4.1 Sensorial multimodality

4.1.1 Introduction

Thanks to the advances in the design of imaging sensors as well as their proliferation, multi-sensors images are now frequently encountered in most fields of image processing. Each resulting multimodal image can be considered as a collection of images of the same scene:

$$\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_P\}, \quad P \geq 2 \tag{4.1}$$

where each $\mathcal{I}_i : E_i \to V_i$ is produced by a different sensor, hence capturing a particular facet of the imaged scene. Although the nature and structure of each image domain E_i could vary from one modality to the other, we restrict in the following to the case where all the modalities share the same domain $E_1 = \cdots = E_P \equiv E$. It implies in particular that all modalities are co-registered. On the other hand, all pixel value sets V_i are not restricted to be the same, and can therefore be of different dimensionality.

Considering all individual modalities at the same time provides a wealth of information and allows a better and more complete description of the scene. This information contained in a multi-source image can be decomposed in two parts: the *redundant* information, *i.e.*, the one that is common to several modalities, and the *complementary* information, which on the contrary is brought by a single modality only. Of course, the definition of each type of information is related to the nature of the sensors composing the multi-source image. For example, consider a multimodal image composed of airborne hyperspectral and LiDAR sensors.

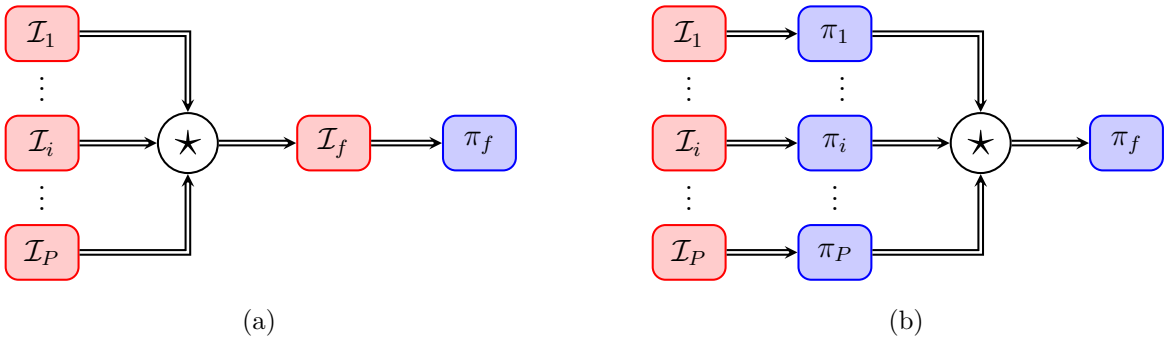


Figure 4.1: Multimodal image segmentation flowcharts, where the fusion operation (represented by the symbol $\star$) occurs at (a) the feature level and (b) the decision level

While the former captures the information related to the spectral response of the materials composing the scene, the latter describes their elevation with respect to some reference height. Considering a scene composed of a concrete building (or any structure above the ground) surrounded by a flat grassy area. In both cases, the properties expressed by each sensor are sufficient to differentiate the structure from the rest of the image (as it is different both in terms of spectral response and height): the information is redundant. On the other hand, suppose now that the structure is a parking lot, such that its height and this of the surrounding area are the same. In such case, it is not be possible to discriminate between the two using the LiDAR information only, and the information brought by the hyperspectral sensor now appears as complementary with respect to the LiDAR modality.

However, the design of adapted tools to process sensorial multimodal images remains a challenge, notably due to the diverse physical meanings and contents of images produced by all possible imaging sensors. Image segmentation is a particular application that would surely benefit from the development of such multimodal tools. The segmentation of a multimodal image should be enhanced thanks to both the complementarity and redundancy of its modalities that should ensure a more robust and accurate delineation of its regions, in particular when those regions share similar features in one mode but not in the other ones. The use of this information can be integrated at two different stages of the processing chain when performing multimodal image segmentation:

– At the *feature level*. In such case, some features are extracted independently from each modality $\mathcal{I}_i, i = 1, \ldots, P$. These are further combined in order to produce some unified feature map and a fused image $\mathcal{I}_f$ from which the final multimodal segmentation is derived, as illustrated by figure 4.1a. In [156] for instance, a general framework for multimodal image fusion is described based on multiresolution (MR) decompositions. Each modality is decomposed using some MR transformation (such as pyramid or wavelet transforms), which are further all merged to created a single combined MR. This latter is finally inverted to retrieved the fused image, on which classical segmentation algorithms can be applied. A similar idea is considered in [56], where the fusion is performed with independent component analysis (ICA). Using some preliminary segmentation of each modality, ICA coefficients are estimated within each region and later fused and inverted

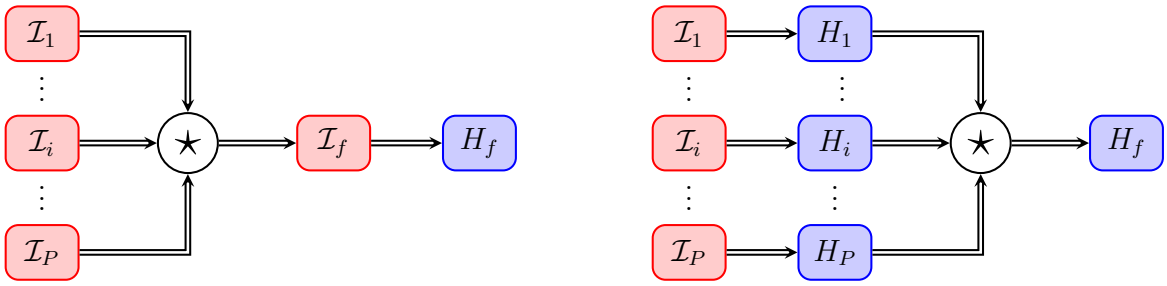


Figure 4.2: Multimodal hierarchical image segmentation flowcharts, where the information fusion (represented by the symbol $\star$) occurs at (a) the feature level and (b) the decision level

to obtain the fused image. Segmentation as input feature is also investigated in [224], where a partition for each modality is first obtained by using its dual tree complex wavelet transform [100]. Features are then modeled for each region by a bi-variate alpha stable distribution and the KL divergence is used to evaluate the similarity between corresponding regions of the modalities. Finally, the fused image is created by merging all regions in the complex wavelet domain prior to inverting it. Note that, for the two latter cases, the segmentation step is only used to provide some regions, considered as features for the fusion process.

– At the *decision level*. In this scenario, each modality $\mathcal{I}_i, i = 1, \ldots, P$ is processed individually and a segmentation map π_f is output independently from the other modalities. These various segmentation are then further combined in order to produce the final multimodal segmentation map π_f, as depicted by figure 4.1b. As for the feature level, several solutions have been proposed to perform the merging of several segmentation maps, ranging from [169] where it is tackled as a geometrical interpolation problem, using distance transforms for each input segmentation in order to find their optimal geometrical average to [225] that makes use of a random walker approach to segment a graph generated based on the homogeneity degree between the input segmentations and [76] where a consensus is reached by ensemble clustering.

Hierarchical structures offer several advantages for various image processing tasks with respect to pixel-based and region-based representations, since they are naturally able to accommodate for the intrinsic multi-scale nature of images, as developed in section 1.3. Image segmentation is one of those tasks, as it reduces to performing a pruning graph cut of the tree representation of the hierarchy. Among all the efficient existing pruning strategies, energy minimization has been widely investigated for hierarchical image segmentation (see chapter 2), as it is highly tunable through an appropriate definition of the energy and thus can be adapted to a wide range of applications.

The representation of multisource images with hierarchical structures seems however to be one step further in terms of difficulty with respect to the multimodal segmentation task. Similarly to the latter, the fusion process to derive a hierarchical representation of a multimodal

image can occur either at the feature level or the decision level, as shown by figure 4.2. The former strategy (figure 4.2a) aims at building a unique hierarchy H_f, that directly encompasses all the specificities of the scene, on a "fused" image $\mathcal{I}_f$ [165]. If the final goal is to derive a multimodal segmentation of the scene, the decision level in that case is eased since classical tools to extract a segmentation from a hierarchy can be applied. However, it may not be easy to make all the modalities cooperate during the construction of the hierarchy, and some features may be "averaged out" by the consensus strategies that have to be adopted during the construction. On the other side, performing the fusion at the decision level implies that several hierarchies $H_i, i = 1, \ldots, P$ have to be somehow combined. In that approach, each hierarchy can capture all the specificities of the modality it is built on, but the fusion decision may become complicated due to the increased number of disagreements that could occurs between the hierarchies.

4.1.2 Objectives of this chapter

Recently, the concept of braids of partitions has been introduced [101, 104] as a potential tool to tackle the fusion of several hierarchies of partitions, and we sketched in [197] how this notion could be adapted in practice to achieve the segmentation of remotely sensed multimodal images. Those preliminary results are extended in this present chapter to define a complete methodology for the hierarchical segmentation of multimodal images based on this concept of braids of partitions.

More specifically, we propose guidelines to derive a braid structure from various cuts of independent hierarchies. Following, we introduce a novel energy function designed to operate on multimodal images in order to extract from the braid an optimal segmentation following the hierarchical energy minimization procedure. The methodology is subsequently tested and validated on two unrelated multisource images featuring different characteristics.

4.2 Energetic optimization on lattices

In chapter 2, we presented the hierarchical energy minimization framework as it was introduced in the work of Guigues [86, 87]. In particular, it was restricted to separable energies only, *i.e.*, when the energy of the partition $\pi = \{\mathcal{R} \subseteq E\}$ is expressed as the sum of the energies of its regions:

$$\mathcal{E}(\pi) = \sum_{\mathcal{R} \in \pi} \mathcal{E}(\mathcal{R}) \ . \tag{4.2}$$

While this formalism encompasses most of the classical energies encountered in the literature, some other composition laws than the sum can be investigated. In particular, we proposed in chapter 2 to define the energy of a partition π as the supremum of the energies of the regions composing it, *i.e*, when $\sum$ is replaced by $\bigvee$ in equation (4.2), and we checked that all theoretical results (such as the minimization procedure and ordering of the optimal cuts for a parametrized energy) were still holding. This therefore raises the question: *which are*

the most general conditions on the definition of the energy function to preserve the theoretical results holding for separable and max-composed energies? This question has been recently answered by Kiran (see for instance [101] and [103]), whose major results are summarized in the following (the reader is referred to Kiran's PhD manuscript [101] for the proofs of all following results).

4.2.1 Energetic ordering and optimization

The starting point of Kiran's approach actually resides in the gigantic cardinality of Π_E, the set of all partitions of the space E, given as the Bell's number $B_{|E|}$ where $|E|$ is the number of elements contained in E. Recall that a 5×5 image can be divided into $B_{25} = 4.6 \times 10^{18}$ different partitions. Assuming that the energy function $\mathcal{E}$ associated with the partition ranges between 0 and 10^6, it means that each energy value is mapped to 4.6×10^{12} different partitions. Finding an "optimal" partition in this setup rather looks like an ill-posed problem.

To alleviate this issue, two solutions stand out:

1. Constraining the space of valid partitions. As discussed in chapter 2, this can be achieved by working on the set of cuts $\Pi_E(H)$ of a hierarchy of partitions H of E. While its cardinality remains impossible to evaluate in practice (Guigues empirically estimated [86, p.157] that a binary hierarchy, *i.e.*, when each node has either two children or none, built over a natural image contains between $1.3^{|\pi_0|}$ and $1.4^{|\pi_0|}$ different cuts, where $|\pi_0|$ is the number of leaves the hierarchy is built on), it is strongly reduced with respect to Π_E
2. Shifting the minimization procedure from the space of the energy values (as the way to evaluate the optimality of a partition is through the value of its energy function) to another space where finding the minimum would be simplified.

While the first solution, working with hierarchies of partitions, is already known, Kiran proposes to investigate the second point. In particular, his base idea is to make the minimization hold on a lattice built on the partitions rather than on the lattice of integer (the space of the energy values). While it is known that Π_E and $\Pi_E(H)$ are two lattices with respect to the refinement ordering $\leq$ (the latter being a sub-lattice of the former), they do not help a lot in this task since the refinement ordering $\leq$ for partitions is not related with any energetic considerations whatsoever. Therefore, a new ordering, defined with respect to the energy function $\mathcal{E}$ must be defined.

4.2.1.1 The energetic ordering $\preceq_{\mathcal{E}}$

In the following, let H be some hierarchy of partitions built over the space E.

Definition 4.1 (Singularity)
Let $\mathcal{E}$ be an energy function as defined by definition (2.1). $\mathcal{E}$ is said to be singular when for every $\mathcal{R} \subseteq E$, $\mathcal{E}(\mathcal{R})$ is either strictly smaller, or strictly greater, than the energy of all the

possible partial partitions $\pi(\mathcal{R})$ *of* $\mathcal{R}$*:*

$$\mathcal{E} \text{ is singular} \Leftrightarrow \forall \mathcal{R} \subseteq E, \mathcal{E}(\mathcal{R}) \left\{ \begin{array}{l} > \bigvee\{\mathcal{E}(\pi(\mathcal{R}))\} \\ < \bigwedge\{\mathcal{E}(\pi(\mathcal{R}))\} \end{array} \right\} \forall \pi(\mathcal{R}) \in \Pi_{\mathcal{R}} \tag{4.3}$$

Note that, in definition (4.1), the partial partition $\pi(\mathcal{R})$ of $\mathcal{R}$ is considered at its broadest sense and not necessarily with respect to the hierarchy H.

Definition 4.2 (Energetic relation $\preceq_{\mathcal{E}}$)
Let $\Pi = \{\pi \in \Pi_E\}$ *be some family of partitions and let* $\pi_i, \pi_j \in \Pi$*.* π_i *is said to be less energetic than* π_j*, and one write* $\pi_i \preceq_{\mathcal{E}} \pi_j$*, when in each region* $\mathcal{R}$ *of* $\pi_i \vee \pi_j$*, the partial partition* $\pi_i(\mathcal{R})$ *of* $\mathcal{R}$ *in* π_i *has a lower energy than the partial partition* $\pi_j(\mathcal{R})$ *of* $\mathcal{R}$ *in* π_j*:*

$$\pi_i \preceq_{\mathcal{E}} \pi_j \Leftrightarrow \forall \mathcal{R} \in \pi_i \vee \pi_j, \mathcal{E}(\pi_i(\mathcal{R})) \leq \mathcal{E}(\pi_j(\mathcal{R})) \tag{4.4}$$

Theorem 4.1
The relation $\preceq_{\mathcal{E}}$ *of definition (4.2) is an ordering for all singular energies* $\mathcal{E}$*, called the energetic ordering, if and only if the family* Π *is the set of cuts* $\Pi_E(H)$ *of a hierarchy of partitions* H*.*

Proof. One can check in [101, pp.30-32] that the relation $\preceq_{\mathcal{E}}$ satisfies the transitivity, reflexivity and anti-symmetry if and only if the compared partitions are cuts of a hierarchy. □

Following the definition of the energetic ordering $\preceq_{\mathcal{E}}$ related to the energy $\mathcal{E}$, one can finally construct a new lattice holding on $\Pi_E(H)$ with the next theorem:

Theorem 4.2
Let $\Pi_E(H)$ *be the set of cuts of a hierarchy of partitions* H*.* $\Pi_E(H)$ *forms a complete lattice for the ordering* $\preceq_{\mathcal{E}}$*. In particular, given a family of cuts* $\pi_i \in \Pi_E(H)$*, the infimum* $\bigwedge_{\mathcal{E}} \pi_i$ *(respectively, supremum* $\bigvee_{\mathcal{E}} \pi_i$*) is obtained by taking the partial partition of lowest (respectively, highest) energy in each region of the refinement supremum* $\bigvee \pi_i$*.*

Proof. The proof is given in [101, p.33]. □

The global infimum of this lattice $(\Pi_E(H), \preceq_{\mathcal{E}})$ is denoted $\pi^{\diamond} = \bigwedge_{\mathcal{E}}\{\pi, \pi \in \Pi_E(H)\}$. It is the unique cut of H that is smaller than all the other cuts with respect to the energetic ordering $\preceq_{\mathcal{E}}$.

4.2.1.2 h-increasingness and optimal cuts

Three different lattices are handled at this stage of the approach developed by Kiran:
- The classical numerical lattice $(\mathbb{R}, \leq)$ for the values of the energy $\mathcal{E}$.
- The refinement lattice $(\Pi_E(H), \leq)$ for the cuts of the hierarchy H.
- The energetic lattice $(\Pi_E(H), \preceq_{\mathcal{E}})$, also holding for the cuts of H.

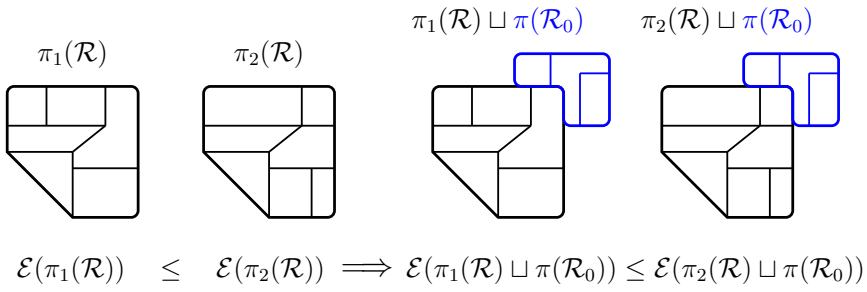


Figure 4.3: Example of a h-increasing energy.

As it was defined by equation (2.14) in chapter 2, the optimal cut $\pi^\star$ of H with respect to the energy $\mathcal{E}$ is defined as the element of $\Pi_E(H)$ whose energy is minimal (*i.e.*, operating on the first of the three lattices mentioned above):

$$\pi^\star = \underset{\pi \in \Pi_E(H)}{\operatorname{argmin}}\ \mathcal{E}(\pi) \tag{4.5}$$

On the other hand, $\pi^\diamond$, the global infimum of the third lattice $(\Pi_E(H), \preceq_\mathcal{E})$, is the minimal cut of H with respect to the energetic ordering $\preceq_\mathcal{E}$. But two cuts being ordered for this energetic ordering does not necessarily implies that their energies are also ordered in the same way, *i.e.* $\pi \preceq_\mathcal{E} \pi' \not\Rightarrow \mathcal{E}(\pi) \leq \mathcal{E}(\pi')$. Under which additional condition on $\mathcal{E}$ can we ensure that $\pi^\star \equiv \pi^\diamond$? To answer this question, Kiran introduces the notion of *h-increasingness* (h standing for hierarchical):

Definition 4.3 (h-increasingness)
An energy $\mathcal{E}$ is said to be h-increasing when given any two disjoint regions $\mathcal{R}, \mathcal{R}_0 \in H$, given partial partitions $\pi_1(\mathcal{R})$, $\pi_2(\mathcal{R})$ and $\pi(\mathcal{R}_0)$, then

$$\mathcal{E}(\pi_1(\mathcal{R})) \leq \mathcal{E}(\pi_2(\mathcal{R})) \Rightarrow \mathcal{E}(\pi_1(\mathcal{R}) \sqcup \pi(\mathcal{R}_0)) \leq \mathcal{E}(\pi_2(\mathcal{R}) \sqcup \pi(\mathcal{R}_0)), \tag{4.6}$$

with $\sqcup$ denoting disjoint union (concatenation). If the inequalities are strict in equation (4.6), the energy is said to be strictly h-increasing.

An example of h-increasing energy is depicted by figure 4.3.

Definition 4.4 (climbing energy)
An energy $\mathcal{E}$ which is both singular and h-increasing is said to be climbing.

The h-increasingness property bridges the gap between the energetic lattice $(\Pi_E(H), \preceq_\mathcal{E})$ of the cuts of a hierarchy and the numerical lattice $(\mathbb{R}^+, \leq)$ of their energy values. Kiran then formulates the following theorem:

Theorem 4.3
Let H be a hierarchy of partitions of E, $\Pi_E(H)$ be the set of its cuts and $\mathcal{E}$ be an energy acting

on it. Let $\pi_i, \pi_j \in \Pi_E(H)$ *be two such cuts. Then*

$$\pi_i \preceq_{\mathcal{E}} \pi_j \Rightarrow \mathcal{E}(\pi_i) \leq \mathcal{E}(\pi_j) \tag{4.7}$$

if $\mathcal{E}$ *is h-increasing. In particular:*

$$\pi^\diamond = \bigwedge\nolimits_{\mathcal{E}}\{\pi, \pi \in \Pi_E(H)\} \Rightarrow \mathcal{E}(\pi^\diamond) = \bigwedge\{\mathcal{E}(\pi), \pi \in \Pi_E(H)\} \tag{4.8}$$

$$\Rightarrow \mathcal{E}(\pi^\diamond) = \mathcal{E}(\pi^\star) \tag{4.9}$$

While the converse is false in general (since several cuts can share the same energy), the implication of equation (4.9) becomes an equivalence when the energy $\mathcal{E}$ is strictly h-increasing, which finally allows to conclude on the equality between the two cuts $\pi^\diamond$ and $\pi^\star$:

Theorem 4.4
Let $\mathcal{E}$ *be a strict h-increasing energy. Then, for a hierarchy of partitions* H*, the optimal cut* $\pi^\star$ *with respect to* $\mathcal{E}$ *and the global infimum of the energetic lattice* $\pi^\diamond$ *coincide:*

$$\underset{\pi \in \Pi_E(H)}{\operatorname{argmin}}\, \mathcal{E}(\pi) = \pi^\star \equiv \pi^\diamond = \bigwedge\nolimits_{\mathcal{E}}\{\pi, \pi \in \Pi_E(H)\} \tag{4.10}$$

Moreover, this optimal cut is unique.

Proof. The proof is given in [101, p.36] □

The property of h-increasingness for the energy $\mathcal{E}$ allows to extend the dynamic program procedure formulated by equations (2.16) and (2.17) for separable energies, and by equations (2.31) and (2.32) for max-composed energies to the wider family of climbing energies:

Proposition 4.1 (Hierarchical energy minimization)
Let H *be a hierarchy of partitions over some space* E*, and let* $\mathcal{E}$ *be some climbing energy. Then, for each region* $\mathcal{R}$ *whose set of children is* $\mathsf{C}(\mathcal{R})$*, the optimal cut of* $\mathcal{R}$ *is given either by* $\{\mathcal{R}\}$ *itself, or by the concatenation of the optimal cuts* $\pi^\star(r)$ *of all children* $r \in \mathsf{C}(\mathcal{R})$ *of* $\mathcal{R}$*:*

$$\mathcal{E}^\star(\mathcal{R}) = \min\left\{\mathcal{E}(\mathcal{R}), \mathcal{E}\left(\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r)\right)\right\} \tag{4.11}$$

$$\pi^\star(\mathcal{R}) = \begin{cases} \{\mathcal{R}\} & \text{if } \mathcal{E}(\mathcal{R}) \leq \mathcal{E}\left(\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r)\right) \\ \displaystyle\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r) & \text{otherwise} \end{cases} \tag{4.12}$$

Moreover, this optimal cut is unique.

The optimal cut of $\mathcal{R}$ is given by comparing the energy of $\mathcal{R}$ and the energy of the disjoint union of the optimal partial cuts of its children, and by picking the one whose energy is the

smallest. The optimal cut of the whole hierarchy is the one of the root node, and is reached by scanning all nodes in the hierarchy in one ascending pass.

Proposition (4.1) gives the theoretical guarantee that, given a climbing energy function $\mathcal{E}$ and a hierarchy of partitions H, it is possible to find the optimal cut of H with respect to $\mathcal{E}$ by solving a dynamic program for each region of the hierarchy, scanned from the leaves to the root. But how to check that an energy is climbing, *i.e.* both singular and h-increasing at the same time?

The h-increasingness is actually rather easy to find out, as most energy functions whose definition is not too fancy are h-increasing. Let us verify it for separable energies: assume that we have two partial partitions $\pi_1(\mathcal{R})$ and $\pi_2(\mathcal{R})$ of a region $\mathcal{R} \in H$ such that $\mathcal{E}(\pi_1(\mathcal{R})) \leq \mathcal{E}(\pi_2(\mathcal{R}))$. Then, considering $\mathcal{R}_0 \in H$ disjoint of $\mathcal{R}$, and $\pi(\mathcal{R}_0)$ a partial partition of $\mathcal{R}_0$, one can write:

$$\begin{aligned}
\mathcal{E}(\pi_1(\mathcal{R}) \sqcup \pi(\mathcal{R}_0)) &= \mathcal{E}(\pi_1(\mathcal{R})) + \mathcal{E}(\pi(\mathcal{R}_0)) && \text{by separability} \\
&\leq \mathcal{E}(\pi_2(\mathcal{R})) + \mathcal{E}(\pi(\mathcal{R}_0)) && \text{by assumption} \\
&\leq \mathcal{E}(\pi_2(\mathcal{R}) \sqcup \pi(\mathcal{R}_0))
\end{aligned}$$

and hence a separable energy is h-increasing. The exact same reasoning proves that max-composed energies are also h-increasing. As a matter of fact, Kiran showed [101, pp.40-42] that all energies which can be expressed as a Minkowski expression:

$$\mathcal{E}(\pi) = \left(\sum_{\mathcal{R} \in \pi} \mathcal{E}(\mathcal{R})^\alpha \right)^{\frac{1}{\alpha}} \tag{4.13}$$

are h-increasing for every $\alpha \in [-\infty, +\infty]$. This property generalizes at one go the results which where known beforehand for energies composed by the sum ($\alpha = 1$) [87, 172], the supremum ($\alpha = +\infty$) [3, 217] and the infimum ($\alpha = -\infty$) [102], notably. It is worth noting that the property of h-increasingness for an energy function only depends on the definition of the composition rule, and not on the expression of the regional energy.

Checking for the singularity property (4.3) of an energy is a little bit more complicated, and one can wonder how the previous results are affected when the singularity assumption is dropped. The singularity property allows to construct the energetic ordering $\preceq_{\mathcal{E}}$ and subsequent lattice structure on $\Pi_E(H)$ with $\pi^\diamond$ as unique global infimum. If this singularity property is discarded, one can no longer build the energetic lattice. However, the dynamic program (4.11) and (4.12) holds thanks to the h-increasingness property. By dropping the singularity property of $\mathcal{E}$, one can no longer guarantee the uniqueness of the optimal cut, as $\{\mathcal{R}\}$ and $\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r)$ should be equally propagated as solutions in the dynamic program procedure when their energies are equal (which case can never happen with the singularity holding). At the end of the day, one can obtain several different cuts with minimal energy. However, by always selecting either $\{\mathcal{R}\}$ or $\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r)$ in case of equality, one can reintroduce the unicity of the solution, except that the answered question is "find the largest (or smallest) cut that minimizes the energy" and no longer "find the cut that minimizes the energy".

4.2.2 Climbing families of energies

In chapter 2, we saw that energies are often parameterized by a numerical coefficient λ, generally acting as a trade-off between a goodness-of-fit (GOF) term $\mathcal{E}_\phi$ that pushes the optimal cut toward over-segmentation, and a regularization term $\mathcal{E}_\rho$ which favors under-segmentation. While those two competing terms are often added at the regional level (*i.e.*, $\mathcal{E}_\lambda(\mathcal{R}) = \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})$, as it is the case for instance for the Mumford-Shah functional [142], classical MRF formulation [113] as well as the proposed max-composed energies (2.28), (2.29), (2.35) and (2.36)), Kiran considered the most general case $\{\mathcal{E}_\lambda\}_{\lambda\in\mathbb{R}^+}$, where no presupposition is made on the way the energy depends on λ and to the family of optimal cuts $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ it generates.

We noted in chapter 2 that, under some assumptions (namely the regularization term being sub-additive when the energy is separable, or being greater than 0 for max-composed energies), it was possible to order by refinement those optimal cuts, *i.e.*, $\lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2}$. How does this property transpose to the general case $\mathcal{E}_\lambda$ is the question investigated by Kiran [101, 103]. To that purpose, he introduced the property of scale-increasingness:

Definition 4.5 (Scale-increasingness)
An energy $\mathcal{E}_\lambda$ indexed by λ is scale-increasing if for any $\mathcal{R} \in H$, any $\pi(\mathcal{R}) \in \Pi_E(H(\mathcal{R}))$, and any $0 \leq \lambda_1 \leq \lambda_2$,

$$\mathcal{E}_{\lambda_1}(\mathcal{R}) \leq \mathcal{E}_{\lambda_1}(\pi(\mathcal{R})) \Rightarrow \mathcal{E}_{\lambda_2}(\mathcal{R}) \leq \mathcal{E}_{\lambda_2}(\pi(\mathcal{R})). \tag{4.14}$$

With respect to the h-increasing property which compares the same energy $\mathcal{E}$ at two different levels $\pi(\mathcal{R})$ and $\pi(\mathcal{R}) \sqcup \pi(\mathcal{R}_0)$, the scale-increasing property compares two different energies $\mathcal{E}_{\lambda_1}$ and $\mathcal{E}_{\lambda_2}$ at the same level $\pi(\mathcal{R})$. A scale-increasing energy preserves the ordering of the energies between the regions of the hierarchy H and their partial partitions when the parameter λ varies. Combining the scale-increasing and climbing properties leads to climbing families of energies:

Definition 4.6 (Climbing families of energies)
A climbing family of energies $\{\mathcal{E}_\lambda\}_{\lambda\in\mathbb{R}^+}$ is a family of energies which is:
- *scale-increasing with respect to λ;*
- *climbing for any λ.*

Climbing families of energies allow to extend the multiscale minimal cut theorem (2.1) as it was formulated by Guigues [87]:

Theorem 4.5
Let H be a hierarchy of partitions and $\{\mathcal{E}_\lambda\}$ be a climbing family of energies acting on the set of cuts $\Pi_E(H)$ of H. Then, the family $\{\pi^\star_\lambda\}$ of optimal cuts generates a hierarchy $H^\star$ of optimal partitions, i.e.

$$0 \leq \lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2} \tag{4.15}$$

The previous theorem (4.5) is powerful as it allows to transform some hierarchy H into its persistent version $H^\star$ as long as the energy $\mathcal{E}_\lambda$ can be verified to be a climbing family. But similarly to the climbing property, how can the property of scale-increasingness be checked given some family of energies? The following (and last) proposition demonstrates how to easily construct families of energies that naturally satisfy the scale-increasing property.

Proposition 4.2
If the family $\{\mathcal{E}_\lambda\}_{\lambda\in\mathbb{R}^+}$ is viewed as a mapping $\lambda \mapsto \mathcal{E}_\lambda$, and if this mapping is increasing, then the family $\{\mathcal{E}_\lambda\}_{\lambda\in\mathbb{R}^+}$ is scale-increasing.

Proof. The proof, given in [101, p.50], is also reproduced in Appendix A. □

4.3 Braids of partitions

When dealing with multimodal images, one has to make the most of the complementarity between the different modalities while preserving the information that is shared by those modalities. In particular, if a region is salient in two modalities, it will very likely lead to two nodes in their respective hierarchical representations that have the same spatial support. But identifying the nodes that have the same spatial support in several (possibly huge) hierarchies is a greedy task. Moreover, the opposite phenomenon, namely when a region is salient in one mode and not in the other one, also has to be handled.

4.3.1 Definition of a braid

Braids of partitions have been recently introduced in [104] as a potential tool to combine multiple hierarchies and thus answer the two questions previously raised to tackle segmentation of multimodal images. In our previous work [197], we sketched that the braid structure could in fact be of help when facing multimodal data fusion.
Braids of partitions are defined as follows:

Definition 4.7 (Braid of partitions)
A family of partitions $B = \{\pi_i \in \Pi_E\}$ is called a braid of partitions whenever there exists some hierarchy H_m, called monitor hierarchy, such that:

$$\forall\, \pi_i, \pi_j \in B, \pi_i \vee \pi_{j\neq i} \in \Pi_E(H_m)\backslash\{E\} \tag{4.16}$$

In other words, a braid is a family of partitions such that the refinement suprema of any pair of different partitions of the family are hierarchically organized (in the sense that they define cuts of a hierarchy of partitions H_m), even though the partitions composing the braid might not be. For this reason, braids of partitions are more general than hierarchies of partitions: while hierarchies are braids, the converse is not necessarily true. As a matter of fact, a braid $B = \{\pi_i \in \Pi_E\}$ is the most general family of partitions such that its set of

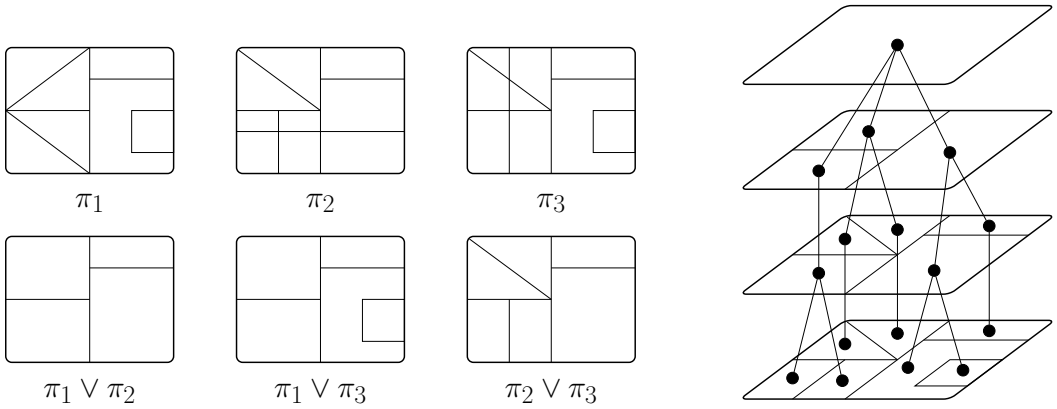
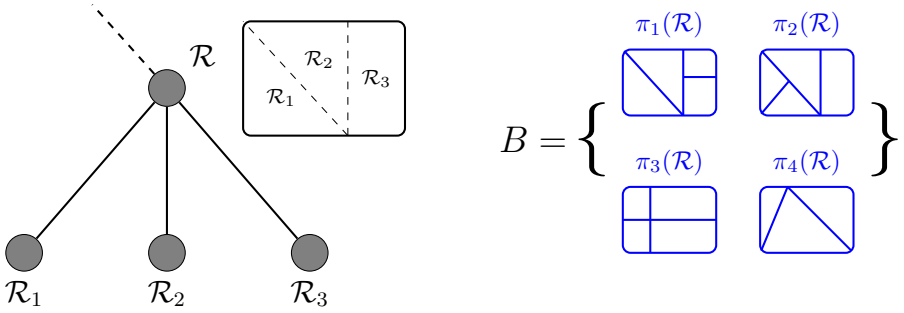


Figure 4.4: Example of braid of partitions $B = \{\pi_1, \pi_2, \pi_3\}$. The hierarchy on the right is a monitor hierarchy of B since all the pairwise refinement suprema $\pi_i \vee \pi_j, i, j \in \{1, 2, 3\}, i \neq j$ define cuts of this hierarchy different from the whole space E.

cuts $\Pi_E(B)$ (still defined as the set of all partitions whose regions belong to B) can replace $\Pi_E(H)$ in the statements of all theorems presented in section 4.2 without invalidating their conclusions. Therefore, all results on the energetic lattice and its global infimum coinciding with the optimal cut with respect to the energy $\mathcal{E}$ remains valid when this energy is operated on a braid B rather than on a hierarchy of partitions H.

It is also worth noting that the refinement supremum of any two partitions must differ from the whole image $\{E\}$ in (4.16). Otherwise, any family of arbitrary partitions would form a braid with $\{E\}$ as a supremum, thus loosing any interesting structure. An example of braid of partitions is displayed by figure 4.4: $B = \{\pi_1, \pi_2, \pi_3\}$ is composed of three partitions which are not comparable by refinement. However, their pairwise refinement suprema are hierarchically related since they are all cuts of the hierarchy on the right, which is by definition a monitor hierarchy of B. Note however that this monitor hierarchy is not unique: one would have the same result by inserting an additional level composed of two regions right below the root of the hierarchy.

The structure of a braid of partitions B, along with its monitor hierarchy H_m, appears to be well suited for the hierarchical representation of multimodal images. As it can be observed in figure 4.4, the monitor hierarchy H_m encodes all regions that are common to at least two different partitions contained in B. Assuming that these partitions originate from different modalities, the monitor hierarchy therefore expresses regions that are salient across the modalities, at various scales. In other word, the monitor hierarchy can be seen as a representation of the redundant information contained in the multimodal image. On the other hand, the family B exhibits the complementary information: all regions contained in B but not in H_m belong to a single modality, and can thus be considered as complementary information. Therefore, the couple B/H_m can be viewed as a hierarchical representation of the multimodal image that relies both on the complementary and redundant information contained in the data.



$$\mathcal{E}^\star(\mathcal{R}) = \min \left\{ \mathcal{E}(\mathcal{R}), \quad \mathcal{E}\left(\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r) \right), \bigwedge_{\pi_i(\mathcal{R}) \in B} \mathcal{E}(\pi_i(\mathcal{R})) \right\}$$

Figure 4.5: Illustration of a step of the dynamic program (4.17) applied to a braid structure: one has to choose between $\{\mathcal{R}\}$, $\bigsqcup \pi^\star(\mathcal{R}_i)$ or any other $\pi_i(\mathcal{R}) \in B$. Note however that $\mathcal{R} \neq E$, otherwise B would not be a braid since $\pi_3(\mathcal{R}) \vee \pi_4(\mathcal{R}) = \mathcal{R}$.

4.3.2 Minimizing an energy function over a braid

Braids of partitions generalize hierarchies of partitions in the sense that the refinement ordering relation between the partitions composing the braid no longer needs to exist. However, as it was shown by [101], braids of partitions are the most general class of families of partitions on which it is possible to construct the energetic ordering $\preceq_\mathcal{E}$ (4.4) on its set of cuts $\Pi_E(B)$, from which derives a lattice structure $(\Pi_E(B), \preceq_\mathcal{E})$.

Therefore, under the same condition of a climbing energy $\mathcal{E}$, the optimal cut of a braid $\pi^\star_B = \operatorname{argmin}_{\pi \in \Pi_E(B)} \mathcal{E}(\pi)$ can be found by solving the dynamic program (4.11) and (4.12) for every region $\mathcal{R}$ of the monitor hierarchy H_m, with however a slight modification due to the fact that, in a braid, any two regions are not necessarily disjoint nor nested:

$$\mathcal{E}^\star(\mathcal{R}) = \min \left\{ \mathcal{E}(\mathcal{R}), \mathcal{E}\left(\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r) \right), \bigwedge_{\pi_i(\mathcal{R}) \in B} \mathcal{E}(\pi_i(\mathcal{R})) \right\} \tag{4.17}$$

$$\pi^\star(\mathcal{R}) = \begin{cases} \{\mathcal{R}\} & \text{if } \mathcal{E}^\star(\mathcal{R}) = \mathcal{E}(\mathcal{R}) \\ \displaystyle\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r) & \text{if } \mathcal{E}^\star(\mathcal{R}) = \mathcal{E}\left(\bigsqcup_{r \in \mathsf{C}(\mathcal{R})} \pi^\star(r) \right) \\ \operatorname{argmin}_{\pi_i(\mathcal{R}) \in B} \mathcal{E}(\pi_i(\mathcal{R})) & \text{otherwise.} \end{cases} \tag{4.18}$$

In addition to comparing the energy of $\mathcal{R} \in H_m$ with respect to the energy of the disjoint union of the optimal cuts of its children, one has also to consider all the other partial partitions $\pi_i(\mathcal{R})$ of $\mathcal{R}$ that can be contained in the braid B, since $\mathcal{R}$ represents the refinement supremum of some regions in the braid, and not those regions themselves. The optimal cut of $\mathcal{R}$ is then given by $\{\mathcal{R}\}$, the disjoint union of the optimal cuts of its children or some other partial

partition of $\mathcal{R}$ contained in the braid, depending on which has the lowest energy. A step of this dynamic program is illustrated by figure 4.5: investigating which is the optimal cut of region $\mathcal{R} \in H_m$, one has to compare:

- The energy $\mathcal{E}(\mathcal{R})$ of $\mathcal{R}$ itself.
- The energy of the disjoint union of the optimal cuts of the children $\mathcal{R}_1, \mathcal{R}_2$ and $\mathcal{R}_3$ of $\mathcal{R}$. Note that $\pi^\star(\mathcal{R}_2)$ is necessarily equal to $\{\mathcal{R}_2\}$ as there is no further partial partition $\pi(\mathcal{R}_2)$ of $\mathcal{R}_2$ in the braid B. $\mathcal{R}_1$ and $\mathcal{R}_3$ admit however some partial partitions in the braid B ($\mathcal{R}_1$ has a partial partition in $\pi_2(\mathcal{R})$ while $\mathcal{R}_3$ has one in $\pi_1(\mathcal{R})$. Their optimal cut can therefore be composed of several regions although $\mathcal{R}_1$ and $\mathcal{R}_3$ are both leaves of the monitor hierarchy H_m.
- The energy of all the other partial partitions which are contained in the braid and not supported by regions of the monitor hierarchy. That is the case for instance for $\pi_3(\mathcal{R})$ and $\pi_4(\mathcal{R})$, which are both partial partitions of $\mathcal{R} \in H_m$ but whose regions do not appear in H_m.

Note that the dynamic program to obtain the optimal cut of the braid $\pi_B^\star$ is conducted on its monitor hierarchy H_m. However, this optimal cut may be composed of regions that do not correspond to any node in the monitor hierarchy. It would be the case in the example depicted by figure 4.5 if $\pi_4(\mathcal{R})$ were for instance chosen to be the optimal cut of $\mathcal{R}$. It is also worth adding that the partial partitions $\pi_i(\mathcal{R})$ of $\mathcal{R} \in H_m$ contained in the braid but not expressed by regions of H_m can be considered as “latent" partial partitions of $\mathcal{R}$. Therefore, the braid optimal cut is obtained by solving the dynamic program (4.17) and (4.18) in a bottom-up manner as well.

4.4 Proposed braid-based hierarchical analysis of multisource images

4.4.1 Generating a braid from multiple hierarchies

As pointed out in [104], the two issues that arise when working with braids of partitions are:

1. Validating that a given family of partitions has a braid structure, that is, condition (4.16) is fulfilled. A possible solution is to explicitly compute all the $\binom{|B|}{2} = \frac{|B|(|B|-1)}{2}$ partitions $\pi_i \vee \pi_{j\neq i}, \pi_i, \pi_j \in B$ with $|B|$ being the number of partitions contained in B, and check that they all define cuts of some hierarchy (in other word, they are all pairwise h-equivalent, as defined directly below).
2. Generating general braids of partitions, that is, finding a set of explicit constraints that must be holding on the space of partitions Π_E to ensure that, given a family $B = \{\pi_i \in \Pi_E\}$, any $\pi_i \in B$ satisfying the imposed constraints is equivalent to B being a braid.

When working with a single hierarchy, it is straightforward to compose a braid since the supremum of two cuts of a hierarchy remains a cut of this hierarchy (as the set of cuts of

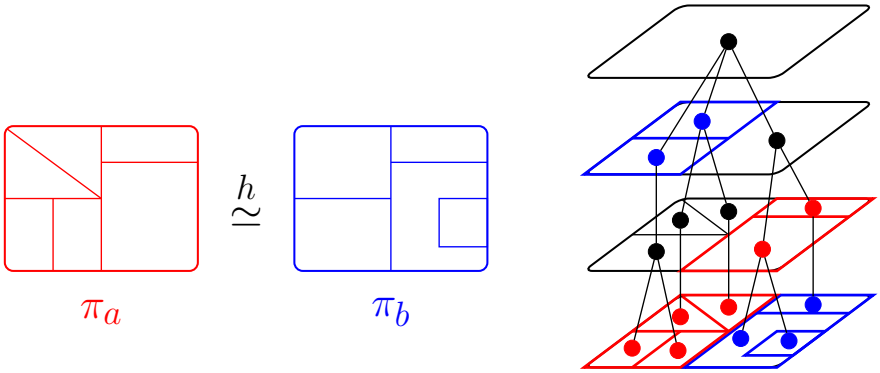


Figure 4.6: Illustration of the h-equivalence relation: π_a, in red, and π_b, in blue, are h-equivalent (left). They define two cuts of the same hierarchy (right).

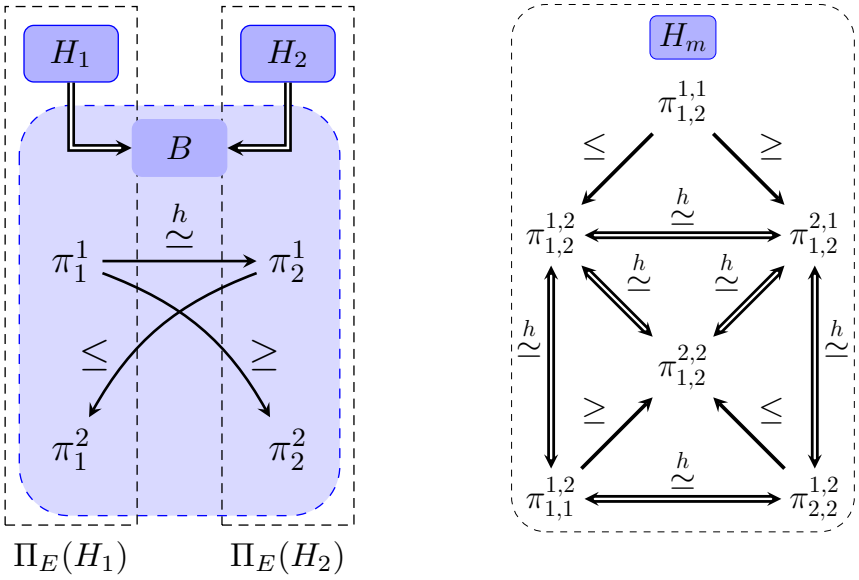
a hierarchy forms a lattice for the refinement ordering). For this reason, any set of cuts $B = \{\pi_i, \pi_i \in \Pi_E(H)\}$ coming from a single hierarchy H is a braid. It also implies in that case that the regions composing the corresponding monitor hierarchy H_m are a subset of the regions composing the initial hierarchy: one does not enrich the hierarchical structure H by composing a braid out of its cuts.

However, this guarantee is lost when one wants to compose a braid from cuts coming from multiple hierarchies H_i, and one has to be careful in that case: all those cuts must be sufficiently related to ensure that all their pairwise refinement suprema are hierarchically organized. Prior to analyzing which constraints must hold on the cuts of various hierarchies to form a braid, we introduce the property of *h-equivalence* (h standing here for *hierarchical*):

Definition 4.8 (h-equivalence)
Two partitions π_a and π_b are said to be h-equivalent, and one notes $\pi_a \overset{h}{\simeq} \pi_b$ if and only if

$$\forall \mathcal{R}_a \in \pi_a, \forall \mathcal{R}_b \in \pi_b, \mathcal{R}_a \cap \mathcal{R}_b \in \{\emptyset, \mathcal{R}_a, \mathcal{R}_b\}. \tag{4.19}$$

In other words, π_a and π_b may not be globally comparable, but they are locally comparable in the sense that some regions of π_a are refined by some regions of π_b, and other regions of π_a are refinements of regions of π_b. For instance, partitions π_a and π_b displayed by figure 4.6 are not globally comparable, but they locally are. Evidently, if two partitions are globally comparable, they are locally comparable as well: $\pi_a \leq \pi_b \Rightarrow \pi_a \overset{h}{\simeq} \pi_b$. Moreover, given a hierarchy H, $\forall \pi_1, \pi_2 \in \Pi_E(H)$, $\pi_1 \overset{h}{\simeq} \pi_2$: all cuts of a hierarchy are h-equivalent, as any two regions of a hierarchy are either disjoint or nested. Conversely, if two partitions are h-equivalent, they define two different cuts of the same hierarchy. $\overset{h}{\simeq}$ is a tolerance relation: it is reflexive and symmetric, but not transitive. To illustrate the lack of transitivity, consider the partitions π_1, π_3 and $\pi_1 \vee \pi_3$ of figure 4.4: $\pi_1 \vee \pi_3$ is h-equivalent with both π_1 and π_3, but π_1 and π_3 are not h-equivalent to each other. Given some hierarchy H and a partition $\pi_* \in \Pi_E$, we denote by $H \overset{h}{\simeq} \pi_*$ the set of cuts of H that are h-equivalent to π_*. Obviously,



(a) (b)

Figure 4.7: (a) Iterative composition of a braid B with cuts from two hierarchies H_1 and H_2 and (b) organization of the cuts of the corresponding monitor hierarchy H_m. For both figures and when cuts are ordered by refinement, the arrow is pointing toward the cut which is the finest of the two.

$H \overset{h}{\simeq} \pi_* \subseteq \Pi_E(H)$ with equality if and only if $\pi_* \in \Pi_E(H)$. Similarly, we denote by $H \leq \pi_*$ the set of cuts of H that are a refinement of π_*.

Now equipped with this h-equivalence relation, let $B = \{\pi_i \in \Pi_E\}$ be a braid, and H_m be a monitor hierarchy of it.

Proposition 4.3
If there exists $\pi_i, \pi_j \in B$ such that $\pi_i \leq \pi_j$, then $\pi_j \in \Pi_E(H_m)$.

Proof. As $\pi_i \leq \pi_j$, it follows that $\pi_i \vee \pi_j = \pi_j$. And from the definition (4.16) of a braid, $\pi_i \vee \pi_j \in \Pi_E(H_m)$, so $\pi_j \in \Pi_E(H_m)$. □

Proposition 4.4
If there exists $\pi_i, \pi_j, \pi_k, \pi_l \in B$ such that $\pi_i \leq \pi_j$ and $\pi_k \leq \pi_l$, then $\pi_j \overset{h}{\simeq} \pi_l$.

Proof. Using the previous proposition (4.3) for both $\pi_i \leq \pi_j$ and $\pi_k \leq \pi_l$, it follows that $\pi_j, \pi_l \in \Pi_E(H_m)$. Using the property of h-equivalence, one concludes that$\pi_j \overset{h}{\simeq} \pi_l$. □

The last proposition has an important consequence in practice: if one wants to compose a braid using two ordered cuts $\pi_i^1, \pi_i^2 \in \Pi_E(H_i), \pi_i^1 \geq \pi_i^2$ coming from two different hierarchies

$H_i, i \in \{1,2\}$, then it is necessary for $B = \{\pi_i^j\}, (i,j) \in \{1,2\} \times \{1,2\}$ to be a braid that $\pi_1^1 \overset{h}{\simeq} \pi_2^1$. Following this, we propose to build a braid using the following iterative procedure:

1. First extract some cut $\pi_1^1 \in \Pi_E(H_1)$. This first cut can be selected arbitrarily.
2. Then choose a cut π_2^1 in the constrained set $H_2 \overset{h}{\simeq} \pi_1^1 \backslash \{E\}$, that is, a cut from H_2 which is h-equivalent to π_1^1 and different from the whole space $\{E\}$.
3. Finally complete by taking a cut in each hierarchy that is a refinement of the cut previously extracted from the other hierarchy, that is $\pi_i^2 \in \Pi_E(H_i), i \in \{1,2\}$ such that $\pi_1^2 \leq \pi_2^1$ and $\pi_2^2 \leq \pi_1^1$.

This procedure is summarized by figure 4.7a (note that, when two cuts are ordered by refinement, the arrow in figure 4.7a is pointing toward the finest cut of the two).

Proposition 4.5
Under this configuration, $B = \{\pi_i^j\}, (i,j) \in \{1,2\} \times \{1,2\}$ has a braid structure with monitor hierarchy H_m whose cuts $\pi_{i,j}^{k,l} = \pi_i^k \vee \pi_j^l$ are organized as displayed by figure 4.7b.

Proof. The proof is given in appendix C. □

While other configurations for the composition of B may also work, *it is the first time that, to the best of our knowledge, guidelines to create a non trivial braid by composing cuts from two independent hierarchies are explicitly provided.* We are, up to now, only able to provide those guidelines and to guarantee the braid structure when at most two cuts are extracted from those two independent hierarchies.

4.4.2 Braid-based multimodal image segmentation

From a conceptual point of view, the braid structure and the subsequent energy minimization procedure conducted on its monitor hierarchy are appealing to perform multimodal segmentation. As a matter of fact, if the braid is composed of partitions extracted from the set of cuts $\Pi_E(H_i)$ of hierarchies H_i constructed on the various modalities $\mathcal{I}_i, i = 1, \ldots, P$, then the monitor hierarchy H_m can be seen as a hierarchical representation containing the salient regions that are common to the various modalities, at all scales. Then, during the energy minimization procedure, the dynamic program has to decide whether a common salient region $\mathcal{R} \in H_m$ should be retained (that is, if $\pi^\star(\mathcal{R}) = \{\mathcal{R}\}$), or replaced either by common regions at a smaller scale ($\pi^\star(\mathcal{R}) = \bigsqcup_{r \in \mathsf{S}(\mathcal{R})} \pi^\star(r)$ with $r \in H_m$ as well) or by a set of regions at a smaller scale, coming from one modality and that fit all the modalities at the same time ($\pi^\star(\mathcal{R}) = \underset{\pi_i(\mathcal{R}) \in B}{\operatorname{argmin}} \mathcal{E}(\pi_i(\mathcal{R}))$).

Therefore, we now propose a methodology to perform multimodal image segmentation based on the concept of braids of partition to fuse the output of several hierarchies. The proposed method is illustrated by the workflow in figure 4.8, detailed step by step in the following. Let $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2\}$ be a multimodal image, assumed to be composed of two modalities $\mathcal{I}_1$ and $\mathcal{I}_2$ having the same spatial support E.

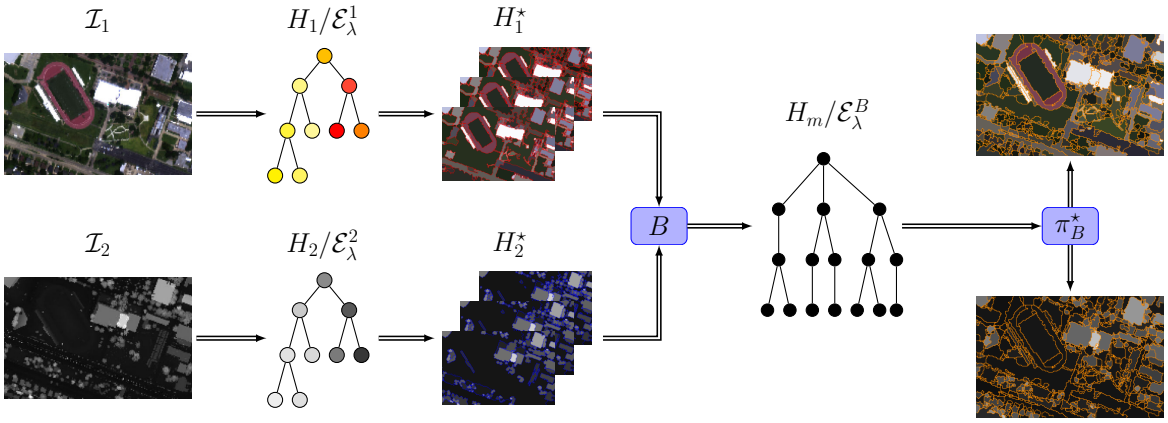


Figure 4.8: Workflow of the proposed braid-based multimodal segmentation methodology.

First, two hierarchies H_1 and H_2 are built on $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. Two energy functions $\mathcal{E}^1_\lambda$ and $\mathcal{E}^2_\lambda$ are defined on their respective hierarchies. The only constraint in practice is that those energy functions must be h-increasing and scale-increasing, in order to be able to transform the hierarchies H_1 and H_2 into their persistent versions $H_1^\star$ and $H_2^\star$ with respect to $\mathcal{E}^1_\lambda$ and $\mathcal{E}^2_\lambda$. As mentioned at the end of section 4.2.1, the singularity property is required to ensure that the optimal cuts coincide with the global infimum of the energetic lattice, for each value of λ. However, if the condition of singularity is dropped, but one consistently chooses either the father or the union of the optimal cuts of the children when their energies are equal during the dynamic program procedure, one can simulate singularity and ensure the uniqueness of the optimal cuts. For segmentation purposes, we propose to define the energy functions as piece-wise constant Mumford-Shah energies [142]:

$$\mathcal{E}^i_\lambda(\pi) = \sum_{R\in\pi}\left(\Xi_i(\mathcal{R}) + \frac{\lambda}{2}|\partial\mathcal{R}|\right) \tag{4.20}$$

where

$$\Xi_i(\mathcal{R}) = \int_{\mathcal{R}} \|\mathcal{I}_i(\mathbf{x}) - \boldsymbol{\mu}_i(\mathcal{R})\|_2^2 dx \tag{4.21}$$

is the GOF term $\mathcal{E}_\phi$ of $\mathcal{E}^i_\lambda$, with $\boldsymbol{\mu}_i(\mathcal{R})$ being the mean value/vector in modality $\mathcal{I}_i$ of the pixel values belonging to region $\mathcal{R}$, and penalizes inhomogeneous regions, thus leading to fine partitions and favoring over-segmentation. The regularization term $\mathcal{E}_\rho$ of $\mathcal{E}^i_\lambda$ is defined as half the length of the region perimeter $|\partial\mathcal{R}|/2$ (note that the coefficient $1/2$ prevent each boundary to account for both the two regions it delimits) and promotes partitions with few region boundaries, therefore favoring under-segmentation on the other hand. The λ coefficient achieves a trade-off to balance the effects of the GOF and regularization terms. The piece-wise constant Mumford-Shah energy function, in addition to being h-increasing (as it is expressed as a separable energy, hence a Minkowski composition (4.13) with $\alpha = 1$) and scale-increasing (following the proposition (4.2) as the functional $\lambda \mapsto \lambda\mathcal{E}_\rho + \mathcal{E}_\phi$ is necessary increasing since $\mathcal{E}_\rho > 0$), is a popular choice when it comes to minimizing some energy function because of its ability to produce consistent segmentations [14, 77]. However, other types of energies could be investigated as well, depending on the underlying application.

Following, the braid B is composed as described previously in subsection 4.4.1 and by figure 4.7a: a first partition $\pi_1^{1\star}$ is extracted arbitrarily from $H_1^\star$, and is used to extract two partitions $\pi_2^{1\star}$ and $\pi_2^{2\star}$ from $H_2^\star$, the first one being h-equivalent and the second one being a refinement of $\pi_1^{1\star}$. A second partition $\pi_1^{2\star}$ is finally extracted from $H_1^\star$ to be a refinement of $\pi_2^{1\star}$. In practice, the sets $H_2^\star \overset{h}{\simeq} \pi_1^{1\star}$, $H_2^\star \leq \pi_1^{1\star}$ and $H_1^\star \leq \pi_2^{1\star}$ may contain several cuts. We propose to define $\pi_2^{1\star}, \pi_1^{2\star}$ and $\pi_2^{2\star}$ as the largest cut of their respective sets, namely $\pi_2^{1\star} = \bigvee\{H_2^\star \overset{h}{\simeq} \pi_1^{1\star} \backslash \{E\}\}$, $\pi_1^{2\star} = \bigvee\{H_1^\star \leq \pi_2^{1\star}\}$ and $\pi_2^{2\star} = \bigvee\{H_2^\star \leq \pi_1^{1\star}\}$. Note that these sets may however be empty, but a workaround to this issue is to built the two hierarchies H_1 and H_2 over the same leaf partition π_0, as it is optimal with respect to $\mathcal{E}_\lambda^i$, for (at least) $\lambda = 0$ (and thus $\pi_0 \in \Pi_E(H_i^\star)$) and is therefore a refinement of all the cuts of $\Pi_E(H_i^\star), i = \{1, 2\}$. Eventually, B is composed of 4 partitions $\{\pi_1^{1\star}, \pi_1^{2\star}, \pi_2^{1\star}, \pi_2^{2\star}\}$ extracted from the two hierarchies $H_1^\star$ and $H_2^\star$, and the braid structure is guaranteed, allowing to construct the monitor hierarchy H_m.

A last energy term $\mathcal{E}_\lambda^B$ is defined as a multimodal piece-wise constant Mumford-Shah energy, relying on both modalities of the multimodal image $\mathcal{I}$:

$$\mathcal{E}_\lambda^B(\pi) = \sum_{\mathcal{R}\in\pi} \left(\max\left(\frac{\Xi_1(\mathcal{R})}{\Xi_1(\mathcal{I}_1)}, \frac{\Xi_2(\mathcal{R})}{\Xi_2(\mathcal{I}_2)} \right) + \frac{\lambda}{2}|\partial\mathcal{R}| \right) \tag{4.22}$$

The GOF term of each region $\mathcal{R}$ is now defined as the maximum with respect to both normalized unimodal GOFs. Here, the maximum is chosen following the idea that a region is optimal if it fits both modalities at the same time. Therefore, a region having a low GOF value with respect to a modality but a high GOF with respect to the other one should be penalized by a high multimodal GOF value. The normalization allows both GOF terms to be in the same dynamical range. $\mathcal{E}_\lambda^B$ is also a h-increasing and scale-increasing energy. Its minimization over H_m and B following the dynamic program (4.17) and (4.18) yields an optimal segmentation $\pi_B^\star$ of $\mathcal{I}$, which should contain salient regions shared by both modalities as well as regions exclusively expressed by $\mathcal{I}_1$ and $\mathcal{I}_2$.

4.4.3 Results assessment

Assessing the consistency of the hierarchical representation of an image in a generic manner is a challenging task, as it greatly depends upon the further application. A common approach is therefore to process the hierarchy accordingly, and appraise the obtained results with respect to the application. The hierarchical model is then declared to be relevant if it leads to proper results. For segmentation purposes, it is widely accepted to compare the output of segmentation algorithms against ground truth segmentation maps, often manually delineated by experts. While this method has been utilized to evaluate hierarchical segmentation results in the case of standard images [10], it is however much more difficult to do so for multimodal images. As a matter of fact, the creation of ground truth data for multimodal images raises several questions since ground truth segmentation maps could either be delineated for each modality and then combined by some means, or directly drawn taking somehow into account all specificities of the multimodal image. In addition, available benchmark multimodal images

are scarce and the manual ground truth delineation is an arduous task. For those reasons, the assessment of hierarchical segmentations for multimodal images is often conducted by visually comparing the multimodal segmentation result against the marginal segmentation outputs (that is, when each modality is processed individually) [165].

To that extend, we propose here to evaluate the efficiency of the braid structure to represent multimodal images by comparing the braid optimal cut $\pi_B^\star$ against the two optimal cuts $\pi_1^\star$ and $\pi_2^\star$ extracted from $H_1^\star$ and $H_2^\star$ and containing the same (or a close) number of regions. This allows a fair visual comparison since all three partitions $\pi_B^\star, \pi_1^\star$ and $\pi_2^\star$ should feature regions of similar scales. In addition, the comparison of partitions with the same (or similar) complexity can be done by evaluating their closeness with respect to the data. For this reason, we compute the average GOF of $\pi_B^\star, \pi_1^\star$ and $\pi_2^\star$ with respect to both modalities $\mathcal{I}_1$ and $\mathcal{I}_2$ as follows:

$$\epsilon(\pi|\mathcal{I}_i) = \frac{1}{|E|}\sum_{\mathcal{R}\in\pi} |\mathcal{R}| \times \Xi_i(\mathcal{R}) \tag{4.23}$$

with $|\mathcal{R}|$ denoting the number of pixels in region $\mathcal{R}$, and $\Xi_i(\mathcal{R})$ is the Mumford-Shah GOF term defined in equation (4.21). Therefore, a consistent braid-based hierarchical representation of the multimodal image should lead to segmentation results competing, for each modality, with its optimal marginal segmentation.

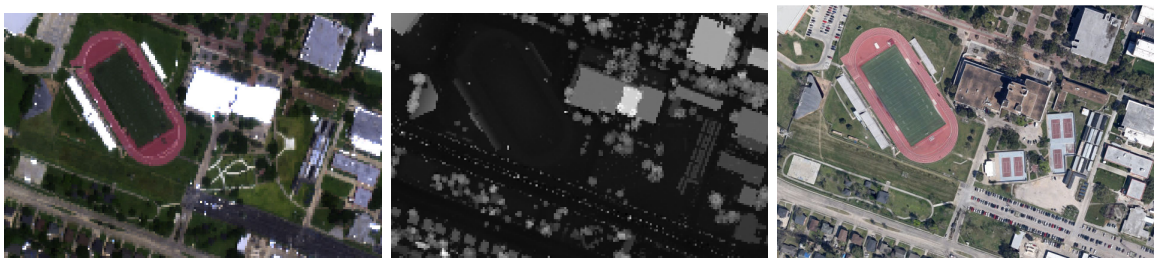
4.5 Experimental validation

In the following, we apply the proposed methodology on two multimodal data sets, each being composed of two co-registered modalities. The first data set is hereafter named Hyperspectral/LiDAR while the second is denoted RGB/depth. For each data set, we first describe its specificities, then we present the experimental set-up used to conduct the multimodal segmentation and we finally display the obtained results.

4.5.1 Hyperspectral/LiDAR data set

4.5.1.1 Description of the data set

The first multimodal data set, described in [61], is composed of a hyperspectral image (HSI) of 144 spectral bands evenly spaced between 380 nm and 1050 nm, and the corresponding LiDAR-derived digital surface model (DSM), with the same ground-sampling distance of 2.5 m. The HSI depicts the spectral reflectance of the scene, *i.e.*, the way the ground has interacted with the incident light. Since each material has an intrinsic reflectance spectrum, HSIs are widely used to identify the different materials composing the scene [157]. The LiDAR image, on the other hand, portrays the height above ground and therefore gives information about the structure or physical shape of the objects composing the scene. The complementarity between the two modalities lies in the fact that two neighboring objects of interest can either be constituted of the same materials but with different heights, or on the



(a) (b) (c)

Figure 4.9: (a) RGB composition of the hyperspectral image, (b) corresponding LiDAR-derived DSM, (c) very-high resolution RGB image of the same site.

other way around, they can share the same height while not being made of the same materials. The hyperspectral/LiDAR multimodality is expected to help discriminate those cases. Each modality has a 120×185 pixels spatial dimension. Data were acquired over the University of Houston campus. The study site features an urban area with several houses and buildings of various heights and roofs made of different materials, an athletics stadium with a running track and two stands, some parking lots, walkways, roads and some portions of grass and trees. A RGB composition of the HSI is displayed in figure 4.9a, and the corresponding LiDAR-derived DSM is shown in figure 4.9b. It is also shown for visualization purpose a very-high resolution RGB image of the scene in figure 4.9c [1]. Note that the very-high resolution image was acquired in 2014, while the HSI and its DSM were acquired in 2012, hence the few changes between the two images (the tennis courts in the center of Fig. 4.9c, notably).

4.5.1.2 Experimental Set-up

The first step of the braid-based multimodal image representation and segmentation methodology is to build the hierarchical representations of the various modalities, as shown by the workflow of figure 4.8. In practice, we use the binary partition tree representation as it already proved to be very efficient for hierarchical image representation and segmentation purposes (see for instance [172, 207, 217] and chapter 2 of the present manuscript). A critical point here is however to build the hierarchical representation appropriately, as an erroneous BPT representation would certainly leads to incorrect cuts and a poorly constructed braid (in terms of multimodal descriptive accuracy). For the Hyperspectral/LiDAR data set, we define the region model and merging criterion as the mean spectrum and spectral angle for the HS modality, and mean value and Euclidean distance for the DSM image, respectively. Those parameters can be considered as standard when working with those types of remotely sensed images. Moreover, the two BPTs H_1 and H_2 are built on the same leaf partition π_0, which is obtained as the refinement infimum of two mean shift clustering procedures [52] conducted on each modality independently (note that the mean shift is run on the RGB composition of the HSI rather than on the HSI directly). This initial partition π_0 features 545 regions.

1. https://goo.gl/maps/Oy5py

Table 4.1: Number of regions $|\pi|$ and average GOF $\epsilon(\pi|\mathcal{I}_i)$ of optimal partitions $\pi_1^\star, \pi_2^\star, \pi_B^\star$ for the Hyperspectral/LiDAR data set, with respect to both modalities $\mathcal{I}_1$ (LiDAR image) and $\mathcal{I}_2$ (hyperspectral image). Lowest values are in bold.

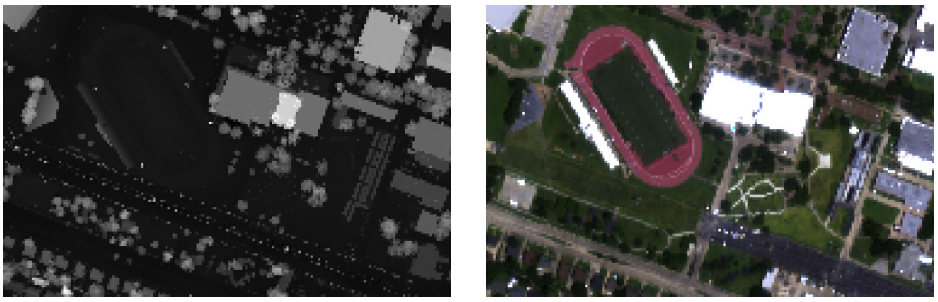
	$\pi_1^\star$	$\pi_2^\star$	$\pi_B^\star$
$\|\pi\|$	325	325	325
$\epsilon(\pi\|\mathcal{I}_1)$	1224.8	3884.8	**994.9**
$\epsilon(\pi\|\mathcal{I}_2)$	145.4	52.5	**48.6**

Constructing the braid B by following the procedure exposed in figure 4.7a raises the question of which hierarchy the first cut should be extracted from. While this is still an open question, we can provide the following rule of thumb, empirically observed during all conducted experiments: the first cut should be extracted from the hierarchy built on the modality whose main regions of interest are the coarsest. Consequently, the first cut is extracted from the BPT built on the LiDAR modality, since it contains less fine details than the HS modality. This first cut, $\pi_1^{1\star}$ contains 150 regions. This number, obtained empirically, roughly corresponds to the number of expected large salient regions in the DSM. It is used to extract $\pi_2^{1\star}$ and $\pi_2^{2\star}$ from $H_2^\star$, which comprise 406 and 414 regions, respectively. Finally, $\pi_1^{2\star}$ is extracted from $H_1^\star$ using $\pi_2^{1\star}$ and contains 495 regions. The four partitions composing B generate $\binom{4}{2} = 6$ cuts of the monitor hierarchy H_m, which is built by re-organizing those cuts in a hierarchical manner. The leaf partition of H_m, denoted π_0^B, is obtained as $\bigwedge\{\pi_i \vee \pi_{j\neq i}, \pi_i, \pi_j \in B\}$. Finally, the minimization of $\mathcal{E}_\lambda^B$ over H_m, following (4.17)), is conducted with λ being empirically set to 5.10^{-5}, and produces an optimal segmentation $\pi_B^\star$ of the braid composed of 325 regions.

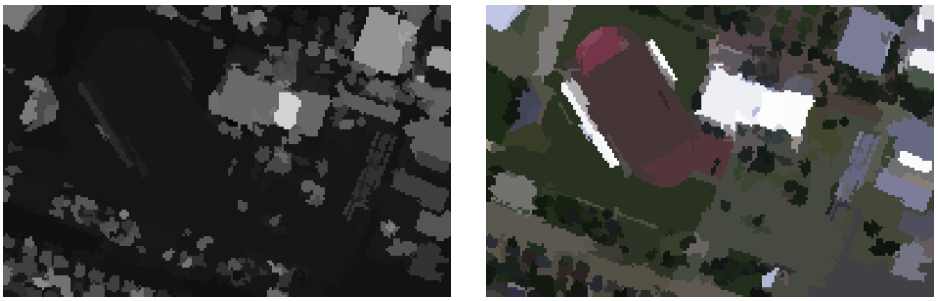
4.5.1.3 Results

Table 4.1 presents the number of regions as well as the average GOF of optimal partitions $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ with respect to both modalities $\mathcal{I}_1$ and $\mathcal{I}_2$. Let us recall that $\mathcal{I}_1$ is the LiDAR-derived DSM modality, and $\mathcal{I}_2$ is the HS modality for this data set. Let us also add that the average GOFs $\epsilon(\pi|\mathcal{I}_i)$ are not absolute values, in the sense that should only be compared values related to the same modality: it is pointless to evaluate the value of $\epsilon(\pi|\mathcal{I}_1)$ against $\epsilon(\pi|\mathcal{I}_2)$ since they are bound to the range of pixel values within $\mathcal{I}_1$ and $\mathcal{I}_2$.

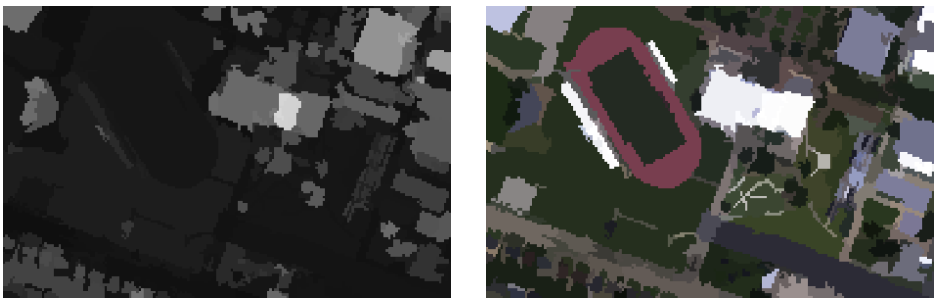
The analysis of table 4.1 demonstrates the effectiveness of the braid structure to make the most of the complementary and redundant information contained within the multimodal data set. As expected, $\pi_1^\star$ and $\pi_2^\star$ score a low average GOF value with respect to their corresponding modality, but a greater average GOF with respect to the complementary modality. On the other hand, $\pi_B^\star$ outperforms both $\pi_1^\star$ and $\pi_2^\star$ with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$. Thus, $\pi_B^\star$, which contains the same number of regions as $\pi_1^\star$ and $\pi_2^\star$, is able to better fit both modalities of the multimodal image at the same time. Remarkably, the average GOF value of $\pi_B^\star$ with respect to $\mathcal{I}_i$ is even lower than the one of $\pi_i^\star$, meaning that the braid structure is able to better delineate



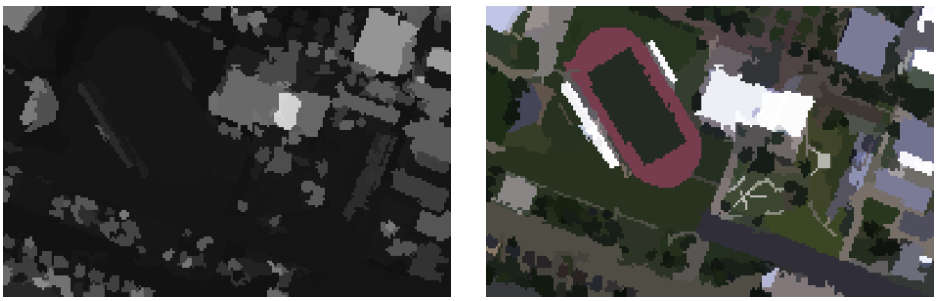
(a) LiDAR modality $\mathcal{I}_1$ (b) HS modality $\mathcal{I}_2$ (RGB composition)



(c) $\pi_1^\star$ over $\mathcal{I}_1$ (d) $\pi_1^\star$ over $\mathcal{I}_2$

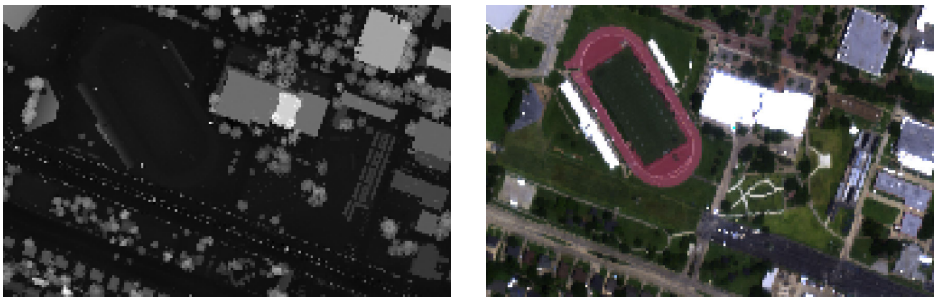


(e) $\pi_2^\star$ over $\mathcal{I}_1$ (f) $\pi_2^\star$ over $\mathcal{I}_2$

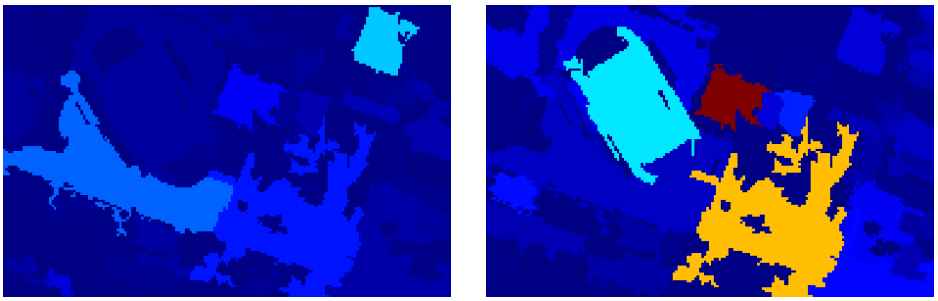


(g) $\pi_B^\star$ over $\mathcal{I}_1$ (h) $\pi_B^\star$ over $\mathcal{I}_2$

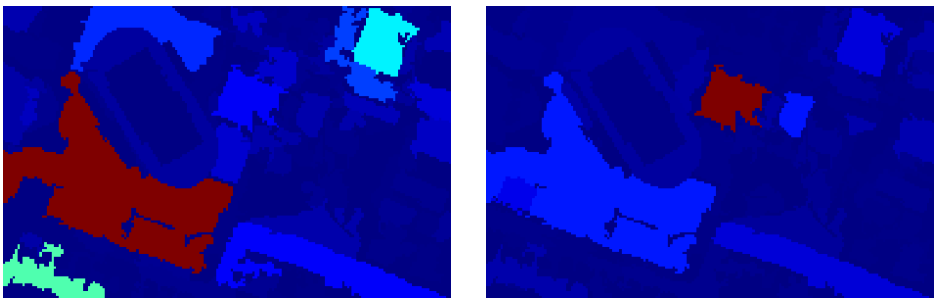
Figure 4.10: Display of optimal partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ represented with their mean height value (over $\mathcal{I}_1$) and mean RGB color (over $\mathcal{I}_2$). All partitions $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ are composed of 325 regions.



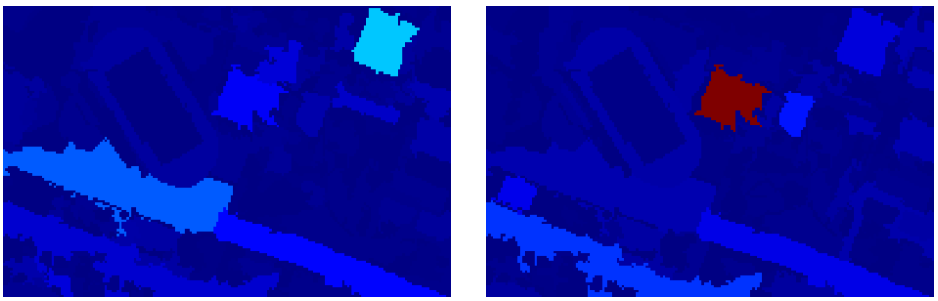
(a) LiDAR modality $\mathcal{I}_1$ (b) HS modality $\mathcal{I}_2$ (RGB composition)



(c) $\epsilon(\pi_1^\star|\mathcal{I}_1)$ (d) $\epsilon(\pi_1^\star|\mathcal{I}_2)$



(e) $\epsilon(\pi_2^\star|\mathcal{I}_1)$ (f) $\epsilon(\pi_2^\star|\mathcal{I}_2)$



(g) $\epsilon(\pi_B^\star|\mathcal{I}_1)$ (h) $\epsilon(\pi_B^\star|\mathcal{I}_2)$

Figure 4.11: Display of the GOF maps associated to partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ with respect to modalities $\mathcal{I}_1$ and $\mathcal{I}_2$. GOF values range from 0 (in blue) to the region-wise maximum over $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ (in red) with respect to the corresponding modality.

the salient regions of $\mathcal{I}_i$ with more accuracy. While this result may seem counterintuitive, it is a perfect illustration of the principle that *the whole is better than the sum of its parts*: the descriptive accuracy and robustness of a multimodal image are increased thanks to the complementarity (for the former) and redundancy (for the latter) of the information contained by each single modality, which are both well exploited by the proposed braid-based framework.

Figure 4.10 shows the two modalities $\mathcal{I}_1$ (figure 4.10a) and $\mathcal{I}_2$ (figure 4.10b), as well as the optimal partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$, represented by their mean LiDAR height (first column, figures 4.10c, 4.10e and 4.10g) and represented by the mean RGB value of the color composition of the HSI (second column, figures 4.10d, 4.10f and 4.10h). Some close-up views for the three optimal partitions $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ are displayed by figure 4.12, with the region boundaries superimposed over the DSM and the RGB composition of the HSI.

The qualitative analysis of figure 4.10 leads to similar conclusions:

- While $\pi_1^\star$ correctly fits $\mathcal{I}_1$ by accurately segmenting all notable regions of the LiDAR modality, such as the various buildings, the trees as well as the houses located on the bottom left corner of the image, it non surprisingly fails at segmenting the regions made of different spectral materials but whose height is similar, such as the running track and the football pitch, or the lawns and roads, as it can be seen in figure 4.10d and figure 4.12b. The reason is straightforward: this cut was extracted from the hierarchy built on height considerations only and thus cannot account for spectrally different regions, provided that they have the same height.
- Contrarily, $\pi_2^\star$ conforms $\mathcal{I}_2$ in the sense that all spectrally salient regions are well preserved. As the explanation is similar to the one provided for $\pi_1^\star$, it can be seen in figures 4.10e and 4.10f that regions which have close spectral signatures but not the same height are generally mis-segmented in $\pi_2^\star$. In particular, several batches of trees are either grouped together, or fused with the neighboring grass (whose spectral response is rather close). Note that for the latter case, despite grass and trees having a close spectral signature, they belong to different semantic classes. This is clearly depicted by figure 4.12c and figure 4.12d, where the group of trees is totally mis-segmented by $\pi_2^\star$.
- When investigating the braid optimal cut $\pi_B^\star$ (see figure 4.10g, figure 4.10h and the two close-up views displayed by figure 4.12e and figure 4.12f), one can see that most erroneous regions of $\pi_1^\star$ and $\pi_2^\star$ are this time correctly delineated. That is notably the case for the running track, the lawns and the roads (with respect to $\pi_1^\star$) or the batches of trees (with respect to $\pi_2^\star$). While this may seem a little bit paradoxical since all $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ all have the same number of regions, a possible explanation is that since those three cuts have a large number of regions (*i.e.*, 325 in that case), $\pi_1^\star$ and $\pi_2^\star$ tend to over-fit their respective modality while $\pi_B^\star$, due to the formulation of the multimodal energy (4.22), is able to better account for "important" details in both modalities.

Finally, figure 4.11 displays the GOF map of each optimal partition $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ (second, third and fourth row, respectively) with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$ (first and second column, respectively). Each map is obtained by assigning to each region $\mathcal{R} \in \pi_1^\star, \pi_2^\star, \pi_B^\star$ its GOF value defined according to the the Mumford-Shah GOF formulation (4.21), so that the values in table 4.1 correspond to the region-wise mean of the corresponding GOF maps.

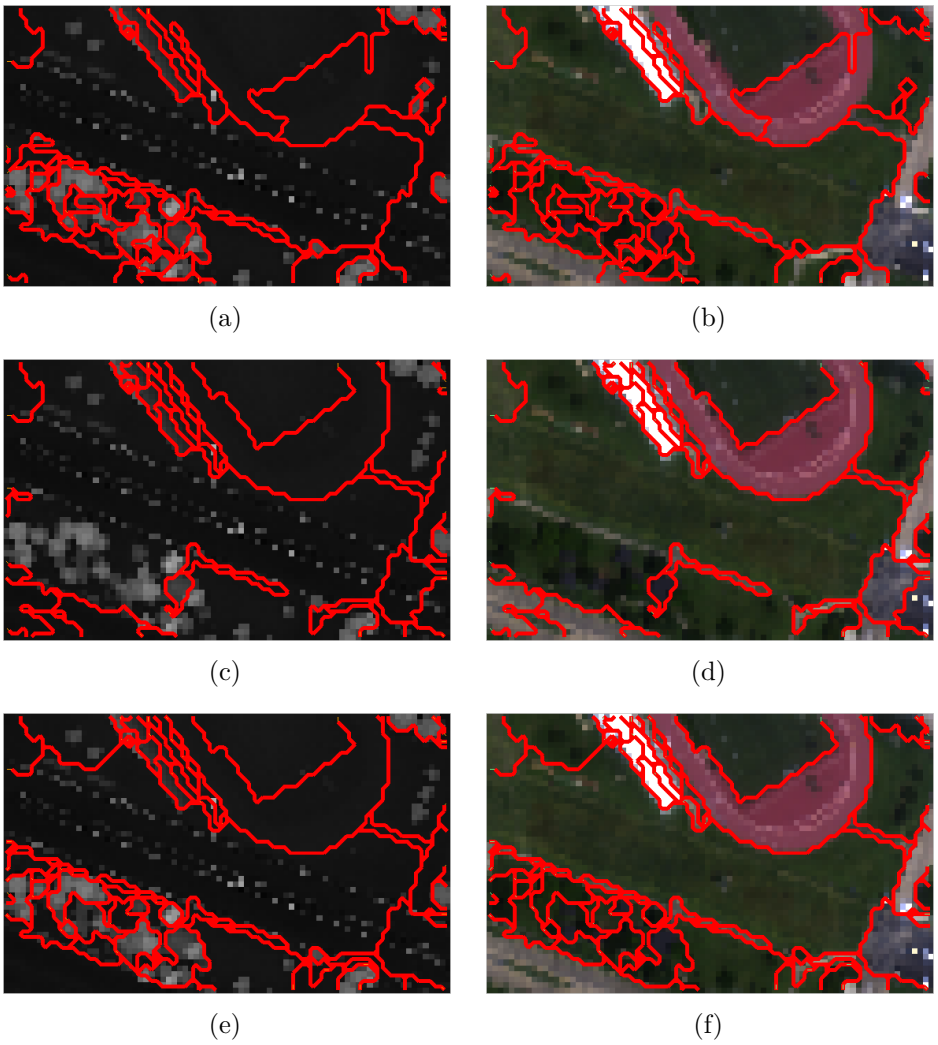


Figure 4.12: Particular of figure 4.10 for $\pi_1^\star$ (top row), $\pi_2^\star$ (middle row) and $\pi_B^\star$ (bottom row) superimposed over $\mathcal{I}_1$ and $\mathcal{I}_2$.

Therefore, each GOF value can be interpreted as the error committed by approximating all pixel values/signatures in $\mathcal{R}$ by their mean (the lower the GOF value, the better). Note that, for visualization purposes, the values have been normalized between 0 (appearing in blue) and the maximum GOF over the regions of $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ with respect to each modality (displayed in red). For that reason, the values associated to the red color for the GOF maps with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$ are not the same (for the LiDAR modality $\mathcal{I}_2$, the maximum GOF is equal to 2.24×10^4 while it is equal to 956 for the hyperspectral modality $\mathcal{I}_2$).

The analysis of figure 4.11 conforms the conclusions drawn from both table 4.1 and figure 4.10. In particular, one can see in figure 4.11c and 4.11d that $\pi_1^\star$ commits globally smaller errors with respect to $\mathcal{I}_1$ than $\mathcal{I}_2$: figure 4.11c is "more blue" than figure 4.11d. As a matter of fact, looking at figure 4.11d, one can see that the highest GOF values correspond to

the athletics stadium and its running track, the large grassy area and roads on the bottom right corner of the image as well as the white building in the middle of the scene. In the first two cases, this is in line with the explanation developed above: while $\pi_1^\star$ is able to properly segment the regions based on their height, it cannot discriminate those which are spectrally different but appear to have the same height. For the last case, the building being red (thus with a maximal GOF value) in all $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$, it suggests that this region was already badly segmented in the initial segmentation map π_0 since it was picked as optimal despite its high GOF value. On the other way around, $\pi_2^\star$ better fits $\mathcal{I}_2$ than $\mathcal{I}_1$ since each region has a lower GOF value with respect to $\mathcal{I}_2$ than to $\mathcal{I}_1$ (except for the building in the center). The largest error in figure 4.11e is associated to the grassy portion on the left of the scene, right below the athletics stadium, and corresponds to the mis-segmentation in $\pi_2^\star$ of the trees and the grass, which have similar spectral responses and thus a relatively low GOF value with respect to $\mathcal{I}_2$, but not with respect to $\mathcal{I}_1$ as their difference in height is important. Finally, the GOF maps associated to $\pi_B^\star$ (figure 4.11g and figure 4.11h), when compared to those of $\pi_1^\star$ and $\pi_2^\star$, validate the potential of the proposed braid-based multimodal segmentation, as the resulting optimal partition $\pi_B^\star$ appears to better fit both modalities at the same time. While this conclusion had already been drawn after the analysis of the global figures in table 4.1, one can also see that given a region $\mathcal{R} \in \pi_B^\star$, its GOF value is smaller than this of the region located at the same place in $\pi_1^\star$ and $\pi_2^\star$ (although the regions may not be the same), both with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$. This confirms that the braid-based methodology is able to delineate more accurate regions than the marginal approaches with respect to both modalities at the same time.

4.5.2 RGB/depth data set

4.5.2.1 Description of the data set

The second considered multimodal image originates from the Middlebury Stereo Dataset [177]. The common usage of this database is the evaluation of two-frame stereo correspondence algorithms (by providing both the stereo images and the ground-truth disparity maps). However, instead of using the left and right frames, which do not seem to be the most suited configuration for segmentation purposes, we rather consider one frame and the associated ground truth depth map, which are displayed by figure 4.13a and figure 4.13b, respectively. In a similar fashion as the Hyperspectral/LiDAR multimodality, the complementarity between the optical and the depth map is expected to benefit the accurate delineation of regions sharing the same properties in one modality but not in the other (for instance, regions appearing with a similar optical color but with different depths with respect to the stereo camera). To reduce the computational burden, each original 2016×2960 image is down-sampled by a factor of 8. Note that the small blurry areas around the umbrellas in figure 4.13b are the result of a basic in-painting method applied on the raw depth map to fill in the missing values [62].

(a) (b)

Figure 4.13: (a) Right view and (b) depth map of the considered stereo data set.

Table 4.2: Number of regions and average GOF of optimal partitions $\pi_1^\star, \pi_2^\star, \pi_B^\star$ for the RGB/depth data set with respect to both modalities $\mathcal{I}_1$ (depth map) and $\mathcal{I}_2$ (RGB image). Lowest values are in bold.

	$\pi_1^\star$	$\pi_2^\star$	$\pi_B^\star$
$\lvert\pi\rvert$	163	158	162
$\epsilon(\pi \lvert \mathcal{I}_1)$	4.2	30.8	**3.9**
$\epsilon(\pi \lvert \mathcal{I}_2)$	51.8	**13.4**	13.9

4.5.2.2 Experimental Set-up

The procedure followed for the RGB/depth data set is identical to the one described above for the hyperspectral/LiDAR data set: the two BPTs are built using region models and merging criteria defined as the mean value and Euclidean distance for each modality. The leaf partition is also identical for both hierarchies, again obtained as the refinement infimum of two mean shift procedures (one per modality), and is made of 506 regions.

The braid B is constructed by first picking a cut from the hierarchy $H_1^\star$ built on the depth map (which is again the modality showing less fine details). This cut, composed of 50 regions, steers the extraction of two cuts from $H_2^\star$, containing 271 and 279 regions, respectively. The final cut is selected from $H_1^\star$ and comprises 417 regions. The construction of the monitor hierarchy H_m is done in the exact same fashion as the previous data set. The multimodal energy $\mathcal{E}_\lambda^B$ is this time minimized with $\lambda = 2.5.10^{-5}$ and leads to the braid-based segmentation $\pi_B^\star$ composed of 162 optimal regions.

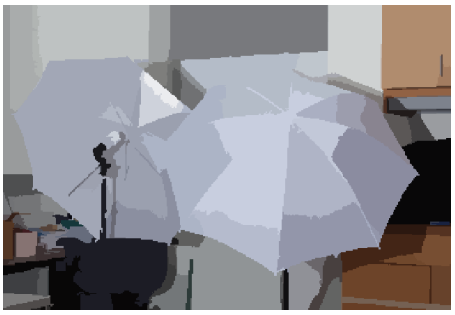
(a) depth map $\mathcal{I}_1$

(b) RGB modality $\mathcal{I}_2$

(c) $\pi_1^\star$ over $\mathcal{I}_1$

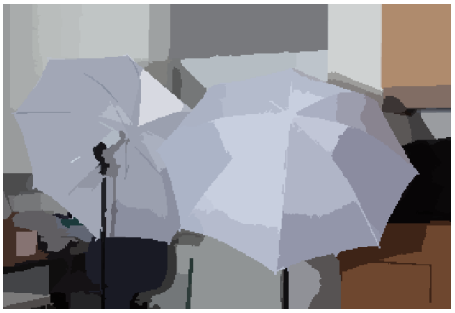
(d) $\pi_1^\star$ over $\mathcal{I}_2$

(e) $\pi_2^\star$ over $\mathcal{I}_1$



(f) $\pi_2^\star$ over $\mathcal{I}_2$

(g) $\pi_B^\star$ over $\mathcal{I}_1$



(h) $\pi_B^\star$ over $\mathcal{I}_2$

Figure 4.14: Display of optimal partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ represented with their mean depth value (over $\mathcal{I}_1$) and mean RGB color (over $\mathcal{I}_2$). Partitions $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ are composed of 163, 158 and 162 regions, respectively.

(a) depth map $\mathcal{I}_1$ (b) RGB modality $\mathcal{I}_2$

(c) $\epsilon(\pi_1^\star|\mathcal{I}_1)$ (d) $\epsilon(\pi_1^\star|\mathcal{I}_2)$

(e) $\epsilon(\pi_2^\star|\mathcal{I}_1)$ (f) $\epsilon(\pi_2^\star|\mathcal{I}_2)$

(g) $\epsilon(\pi_B^\star|\mathcal{I}_1)$ (h) $\epsilon(\pi_B^\star|\mathcal{I}_2)$

Figure 4.15: Display of the GOF maps associated to partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ with respect to modalities $\mathcal{I}_1$ and $\mathcal{I}_2$. GOF values range from 0 (in blue) to the region-wise maximum over $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ (in red) with respect to the corresponding modality.

4.5.2.3 Results

Table 4.2 presents the number of regions as well as the average GOF of optimal partitions $\pi_1^\star, \pi_2^\star$ and $\pi_B^\star$ with respect to both modalities $\mathcal{I}_1$ and $\mathcal{I}_2$. For the RGB/depth data set, we recall that $\mathcal{I}_1$ corresponds to the depth map while $\mathcal{I}_2$ is one frame of the stereo image.

The observations that arise when analyzing table 4.2 are similar to those of table 4.1: each optimal partition $\pi_1^\star$ and $\pi_2^\star$ scores a low average GOF value with respect to its own modality and an increased error with respect to the other one. On the other hand, the braid optimal cut $\pi_B^\star$ outperforms the depth optimal cut $\pi_1^\star$ with respect to $\mathcal{I}_1$ and achieves a comparable value on $\mathcal{I}_2$ with respect to $\pi_2^\star$ (13.9 for $\pi_B^\star$ against 13.4 for $\pi_2^\star$). Let us add that again, the number of regions is similar for $\pi_1^\star$ (163 regions), $\pi_2^\star$ (158 regions) and $\pi_B^\star$ (162 regions), which confirms that the result of the braid-based segmentation is able to better fit both modalities at the same time.

In a same fashion than the Hyperspectral/LiDAR data set, figure 4.14 displays the obtained segmentation maps. The depth map $\mathcal{I}_1$ and RGB image $\mathcal{I}_2$ are shown by figure 4.14a and figure 4.14b, respectively. The optimal partitions $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ are presented by their mean depth value on the first column (figures 4.14c, 4.14e and 4.14g) and by their mean RGB value on the second column (figures 4.14d, 4.14f and 4.14h).

When looking at the specificities of $\mathcal{I}_1$ and $\mathcal{I}_2$, and provided that the conclusions for the RGB/depth data set are in line with those of the Hyperspectral/LiDAR data set, one would expect $\pi_1^\star$ to mis-segment the regions of the background which have the same depth but not the same color and foresee that $\pi_2^\star$ would struggle to properly distinguish the two umbrellas between them and against the background wall, which all have a similar whitish value. As a matter of fact, it can be seen in figure 4.14d that the bottom right corner of the image, which features an half-shaded drawer, is inaccurately segmented, as well as the various objects on top of the desk located on the bottom left corner of the image. Likewise, it can be observed on figures 4.14e and 4.14f that in $\pi_2^\star$, the most forward umbrella has parts which are either confused with the wall behind or with the second umbrella. On the other way around, as shown by figures 4.14g and 4.14h, those regions are well segmented on the braid optimal cut $\pi_B^\star$, confirming again that both modalities have collaborated within the braid framework to design a more accurate segmentation map with respect to the multimodal image.

Eventually, following the results presented for the Hyperspectral/LiDAR data set, figure 4.15 displays the GOF map of each optimal partition $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ (second, third and fourth row, respectively) with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$ (first and second column, respectively). Again, the GOF values have been normalized have been normalized between 0 (in blue) and the maximum GOF over the regions of $\pi_1^\star$, $\pi_2^\star$ and $\pi_B^\star$ with respect to each modality (in red). For the RGB/depth data set, the maximum GOF value with respect to the depth map $\mathcal{I}_1$ is 271, while it is equal to 512 for the RGB modality $\mathcal{I}_2$.

In a similar manner than for the previous Hyperspectral/LiDAR data set, all the observations that arose through the analysis of table 4.2 and figure 4.14 for the RGB/depth data set reverberate on figure 4.15. In particular, one can see in figures 4.15c and 4.15d that $\pi_1^\star$ is

able to represent $\mathcal{I}_1$ with a very low error as the corresponding GOF map appears as almost entirely deep blue, while committing much more errors with respect to $\mathcal{I}_2$, as the background wall, the shaded drawer on the right side as well as parts of the umbrellas have a large GOF value. All those areas indeed have the same depth with respect to the imaging sensor, but not the same color values. Contrarily, $\pi_2^\star$ properly segments $\mathcal{I}_2$ based on color considerations (figure 4.15f), but is unable to accurately distinguish regions which do not have the same depth while having close colors, such as the two umbrellas as well as the background wall. Those regions thus appear with a higher GOF value in figure 4.15e. The optimal braid segmentation $\pi_B^\star$ on the other hand, has low GOF values for all regions, with respect to both $\mathcal{I}_1$ and $\mathcal{I}_2$. It once more corroborates the conclusion that the proposed braid-based methodology is able to make the most both of the redundant and complementary information which are contained in the two modalities, in order to derived a more accurate segmentation of the multimodal image.

4.6 Conclusion

In this final chapter, we focused on the sensorial multimodality, that is, when several images of the same scene are acquired with different sensors. Each sensor featuring its own characteristics, the resulting multisource image gains in descriptive capacities and accuracy. While several typical image processing operations should benefit from this increased amount of information, handling and processing a multisource image in a generic fashion is a real challenge due to the large number of potential multimodalities which can be encountered, engendered by the huge diversity of imaging sensors.

Among all image processing tasks, we investigated the segmentation process, which aims at finding a partition of the image such that each region of the partition is meaningful with respect to the application. Often, this meaning is expressed in terms of semantic, as one expects from the regions of the partition to correspond to real, visually interpretable, regions of the scene. The major downside of the segmentation task is that it is an ill-posed problem: a single image does not admit a single acceptable segmentation, as it can be partitioned at different levels of details. In order to tackle this intrinsic multiscale natures, hierarchies of partitions have been propsed as a successful solution for image representation. Based on the hierarchy, which is built once and regardless of the application, the image segmentation problem amounts at finding a suitable cut (*i.e.*, a partition of the image such that all regions belong to the hierarchy) with respect to the underlying application. This search for a relevant cut can be in practice efficiently carried out by an energy minimization procedure over the space of cuts of the hierarchy.

While the fusion of multiple hierarchies could be a potential solution to the multisource image segmentation issue, it remains a challenge in practice. Recently, braids of partitions have been proposed as a generalization of hierarchies of partitions, since regions belonging to a braid no longer need to be either disjoint or nested. Braids of partitions come along with an associated hierarchy, called the monitor hierarchy. Searching for the optimal cut of a braid reduces to finding the optimal cut of its monitor hierarchy following a slightly different procedure than the common one.

It was conjectured in the seminal work of Kiran [101, 104] that the energy minimization framework over the braid structure could be of use for multivariate optimization. Based on those tools, we derived in this chapter a novel methodology to perform the segmentation of multimodal images. We identified two critical points in this methodology, namely the practical construction of the braid and the definition of the subsequent energy function to be minimized on it. Therefore, we presented guidelines to ensure the composition of several cuts, extracted from hierarchies independently built over each modality, to demonstrate a braid structure. In addition, we proposed a multimodal energy as an adaption of the so-called piece-wise constant Mumford-Shah energy, in order to generate partitions whose regions achieve a trade-off between goodness-of-fit (a high descriptive accuracy) and simplicity.

The proposed methodology was successfully investigated on two different and unrelated multisource data sets. The first one, arising in the remote sensing field, was composed of a hyperspectral image and its corresponding co-registered LiDAR-derived digital surface model. The second, related to stereo vision, comprised a color image, namely the right view of a stereo image, and the associated depth map. In both cases, the obtained results demonstrated, quantitatively and qualitatively, the ability of the proposed approach to produce a segmentation that not only retains salient regions shared by both modalities, but also regions appearing in only one modality of the multimodal image, outperforming the typical marginal segmentation results (considering only one modality independently of the other). We also stress again that the proposed methodology goes even beyond segmentation: building a full multimodal hierarchy, it could be further used in a variety of image processing applications.

From a theoretical aspect, future work will focus on the construction of the braid. As a matter of fact, the proposed method as it currently stands, only allows to extract two different cuts from two independent hierarchies, constraining the set of possible multisource images which can be investigated. Incorporating a higher number of cuts from more hierarchies while maintaining the braid structure appears as a clear line of research. From a practical point of view, the definition of new types of multimodal energies to achieve other applications than segmentation, and their application on different types of multimodalities will also be investigated.

Conclusion

This thesis has been devoted to the study of multimodality and hierarchical representations. The multimodality phenomenon occurs more and more frequently in image processing, and while its benefits for many practical applications is not questioned, its generic handling and exploitation however raises several challenges. Hierarchical representations on the other hand are known to be a powerful tool, as allow to capture the intrinsic multiscale nature of images. Hierarchical representations, and their subsequent processing, have shown to be a valuable tool for several image processing tasks such as image segmentation, image filtering, object detection, and so on.

The objective of this thesis has been the extension of hierarchical representations to multimodal images, in order to better exploit the information brought by the multimodality and design more efficient image processing techniques. The integration of the multimodal information within the hierarchical representation being subject to the nature of the multimodality, and the further design of adapted hierarchical processing techniques being driven by the underlying application, this extension had to be articulated around the quadruplet *multimodality/hierarchical representation/hierarchical processing/application*.

Therefore, we presented in a first instance each element of this quadruplet separately. In particular, we proposed a formal definition to characterize multimodal images. We introduced hierarchical representations from a conceptual point of view, and presented some common instances, focusing more especially on the binary partition tree representation. Finally, we showed in a practical scenario how to properly operate the hierarchical representation of a classical image and its processing in order to achieve a given application. Equipped with those tools, we then turned our attention to multimodal images, focusing in particular on multimodalities frequently occurring in the remote sensing.

Spectral-spatial multimodality

The first multimodality studied in this thesis was the spectral-spatial multimodality. More particularly, we turned our attention to hyperspectral images, as they feature both the information related to the spatial structures and the spectral constituents of the scene. The joint use of the spectral and spatial information has already been widely studied for several typical hyperspectral processing such as hyperspectral classification as well as spectral unmixing. For the latter, the use of spatial information to improve the unmixing performances has been already investigated, but the unmixing procedure has always been conducted over the whole image.

Here, we proposed to investigate the opposite direction, namely to perform the spectral unmixing on local regions in the image. More specifically, we proposed to derive a segmentation (a spatial processing in essence) of the hyperspectral image being adapted for this local spectral

unmixing approach (being a spectral processing). Using a hierarchical representation of the hyperspectral image appeared as the natural solution to the possible varying scale under which the image should be unmixed. Working with the binary partition tree representation, we therefore proposed two novel region models and their associated merging criteria, relying on the results of the local unmixing over the region (namely the local endmembers and their corresponding abundances). The final goal being the derivation of a segmentation (that is, a cut of the hierarchical representation) being optimal with respect to the spectral unmixing, we formulated the optimality criterion as an energy minimization procedure, and proposed some suited energy functions, finally leading to extract from the hierarchical representation a segmentation with minimal reconstruction error.

Among the major perspectives of this work on the use of spectral-spatial multimodality to derive some optimal segmentation with respect to the spectral unmixing is the subsequent analysis and processing of all the generated endmembers, each being locally optimal. We believe in particular that the whole set of endmembers should actually reduce to several slight different instances of a limited number of spectral materials. The proposed method could therefore be a way to reveal the spectral variability of the endmembers. The question on how to finally use these endmembers to finish the unmixing of the hyperspectral image also remains an open question so far.

Temporal multimodality

The temporal multimodality was then investigated, occurring when several images of the scene are obtained at different acquisition dates. The comparison between consecutive images of the multi-temporal sequence reveals which parts of the scene are changing with time. This multimodality is notably of use to perform object tracking. While object tracking is a mature application of computer vision for traditional video sequences, the case of hyperspectral sequences remains largely untreated, as the extension of classical object tracking methods to high dimensional hyperspectral images is challenging, and benchmark hyperspectral video sequences to validated new methods are scarce.

Therefore, we developed a new methodology to perform object tracking for hyperspectral video sequence. Handling the high dimensionality of the hyperspectral frames was embedded in their hierarchical representation, and the object tracking was handled as a sequential object detection process with the hierarchical representation as support. Designed to be a generic methodology, we subsequently tuned it appropriately to perform the tracking of chemical gas plumes in long-wave infrared hyperspectral video sequences, and compared it with two state-of-the-art methods.

Up to now, the proposed method is only able to handle a single moving object. The main perspective of the methodological contribution is the relaxation of this assumption, in order to extend the tracking to multiple moving objects. Related to the chemical plume tracking application on the other hand, the proposed method, as well as the two state-of-the-art methods it was compared with, only provide information related to the position of the plume

in the sequence. The information related to the concentration of the plume (which could be provided by a spectral unmixing operation) could be included in a future step.

Sensorial multimodality

The sensorial multimodality was finally considered in this manuscript. By acquiring images with multiple different sensors, each capturing a particular aspect of the scene, one enhances the descriptive capacity and accuracy of the resulting multisource image. This wealth of information should be beneficial for several image processing tasks, and we decided to focus in particular on the image segmentation application.

Hierarchical representations are a convenient tool to achieve image segmentation, as they can naturally provide various level of details in the resulting segmentation map. Building one hierarchical representation per modality raises however the question on their further fusion. Recently, braids of partitions were proposed as a generalization of hierarchical representations, as the regions contained in the braid no longer need to be either disjoint or nested. They were also conjectured to be a potential solution to the issue on the fusion of multiple hierarchies of partitions. Therefore, we proposed a way to implement braids of partition in the scenario being the segmentation of multisource images. The practical construction of a braid being bounded by its theoretical properties, we derived some guidelines to extract cuts from two independent hierarchies and guarantee the braid structure. Using an energetic framework, the final segmentation of the multisource image was finally obtained as the optimal cut, for a proposed multimodal energy, of the braid structure and its associated monitor hierarchy. The proposed methodology was investigated on two multisource images featuring different characteristics, and the resulting braid optimal cut was found to outperform in terms of average goodness-of-fit the marginal segmentations in both cases.

As it currently stands, the proposed method only allows to extract two different cuts from two independent hierarchies for the construction of the braid, constraining the set of possible multisource images which can be investigated. Incorporating a higher number of cuts from more hierarchies while maintaining the braid structure appears as a clear line of research.

List of publications

International journal

- ⋆ **G. Tochon**, J-B. Féret, S. Valero, R. E. Martin, D. E. Knapp, P. Salembier, J. Chanussot, G. P. Asner. On the use of binary partition trees for the tree crown segmentation of tropical rainforest hyperspectral images. *Remote Sensing of Environment*, 159:318-331, 2015.
- ⋆ M. A. Veganzones, **G. Tochon**, M. Dalla Mura, A. J. Plaza, J. Chanussot. Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation. *IEEE Transactions on Image Processing*, 23(8):3574-3589, 2014.

National journal

- ⋆ **G. Tochon**, J-B. Féret, S. Valero, R. E. Martin, R. Tupayachi, J. Chanussot, P. Salembier, G. P. Asner. Segmentation hyperspectrale de forêts tropicales par arbres de partition binaires. *Revue française de photogrammétrie et de télédétection*, 202:55-65, 2013.

International conference

- ⋆ **G. Tochon**, M. Dalla Mura, and J. Chanussot. Segmentation of multimodal images based on hierarchies of partitions. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 241-252. Springer, 2015.
- ⋆ M. A. Veganzones, M. Dalla Mura, **G. Tochon**, and J. Chanussot. Binary partition tree-based spectral-spatial permutation ordering. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 434-445. Springer, 2015.
- ⋆ **G. Tochon**, J. Chanussot, J. Gilles, M. Dalla Mura, J-M Chang, and A. L. Bertozzi. Gas plume detection and tracking in hyperspectral video sequences using binary partition trees. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.
- ⋆ L. Drumetz, M. A. Veganzones, R. Marrero, **G. Tochon**, M. Dalla Mura, A. J. Plaza, and J. Chanussot. Binary partition tree-based local spectral unmixing. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.
- ⋆ M. A. Veganzones, L. Drumetz, **G. Tochon**, M. Dalla Mura, A. J. Plaza, J. M. Bioucas-Dias, and J. Chanussot. A new extended linear mixing model to address spectral variability. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.
- ⋆ M.A. Veganzones, **G. Tochon**, M. Dalla Mura, A. Plaza, and J. Chanussot. Hyperspectral image segmentation using a new spectral mixture-based binary partition tree representation. In *2013 20th IEEE International Conference on Image Processing (ICIP)*, 2013.

- ⋆ M. A. Veganzones, **G. Tochon**, M. Dalla Mura, A. J. Plaza, and J. Chanussot. A comparison study between windowing and binary partition trees for hyperspectral image information mining. In *2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2013.
- ⋆ **G. Tochon**, J-B. Féret, R. E. Martin, R. Tupayachi, J. Chanussot, and G. P. Asner. Binary partition tree as a hyperspectral segmentation tool for tropical rainforests. In *2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012.

National conference

- ⋆ **G. Tochon**, M. Dalla Mura, and J. Chanussot. Segmentation hiérarchique d'images multimodales. In *XXVeme colloque GRETSI*, 2015.
- ⋆ **G. Tochon**, M. A. Veganzones, M. Dalla Mura, A. J. Plaza, and J. Chanussot. Segmentation hyperspectrale adaptée au démélangeage spectral au moyen d'un arbre de partition binaire. In *XXIVeme colloque GRETSI*, 2013.

Currently under review

- ⋆ **G. Tochon**, M. Dalla Mura, M. A. Veganzones, and J. Chanussot. Braids of partition for the hierarchical analysis of multimodal images. Submitted to *Pattern Recognition*.
- ⋆ **G. Tochon**, J. Chanussot, M. Dalla Mura, and A. L. Bertozzi. Hierarchical representation of hyperspectral video sequences: application to chemical gas plume tracking. Submitted to *Pattern recognition*.
- ⋆ L. Drumetz, M. A. Veganzones, R. Marrero, **G. Tochon**, M. Dalla Mura, G. Licciardi, C. Jutten, and J. Chanussot. Hyperspectral local intrinsic dimensionality. Submitted to *IEEE Transactions on Geoscience and Remote Sensing*.

Appendix

A Multiscale minimal cut theorem for max-composed energies

Theorem (Multiscale minimal cut for max-composed energies)
Let H be a hierarchy on a set E, and let $\mathcal{E}_\lambda(\pi) = \bigvee_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})$ be an affine max-composed energy such that $\mathcal{E}_\rho(\mathcal{R}) \geq 0\ \forall \mathcal{R} \in H$. Then the family of optimal cuts $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ can be ordered by refinement, i.e.:

$$\forall \lambda_1, \lambda_2, 0 \leq \lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2} \tag{A.1}$$

Proof. The proof is adapted from [101, 103] and is organized in three steps: first we prove that the mapping $\lambda \mapsto \mathcal{E}_\lambda$ is increasing. Then, we demonstrate that this monotonicity behavior implies the scale-increasingness of energy $\mathcal{E}_\lambda$. We finally show that the optimal cuts $\{\pi^\star_\lambda\}$ being ordered by refinement derives from the scale-increasingness of the energy function.

1. **Monotonicity of the mapping $\lambda \mapsto \mathcal{E}_\lambda$.** Let $\pi = \{\mathcal{R} \subseteq E\}$ be some partition, and consider the mapping
$$\lambda \mapsto \mathcal{E}_\lambda(\pi) = \bigvee_{\mathcal{R}\in\pi} \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R}) \tag{A.2}$$
which is the maximum of several affine functions $\lambda \mapsto \mathcal{E}_\lambda(\mathcal{R}) = \mathcal{E}_\phi(\mathcal{R}) + \lambda\mathcal{E}_\rho(\mathcal{R})$. By assumption, $\mathcal{E}_\rho(\mathcal{R}) \geq 0$ so the affine function $\lambda \mapsto \mathcal{E}_\lambda(\mathcal{R})$ is increasing for any $\mathcal{R}$ with respect to λ, and so is the maximum $\bigvee_{\mathcal{R}\in\pi} \mathcal{E}_\lambda(\mathcal{R})$.
2. **Scale-increasingness of $\mathcal{E}_\lambda$.** Now consider a region $\mathcal{R} \in H$ and $\pi(\mathcal{R})$ being a partial partition of $\mathcal{R}$, and let $0 \leq \lambda_1 \leq \lambda_2$. By increasing monotonicity of $\mathcal{E}_\lambda$, one has $\mathcal{E}_{\lambda_1}(\mathcal{R}) \leq \mathcal{E}_{\lambda_2}(\mathcal{R})$ and $\mathcal{E}_{\lambda_1}(\pi(\mathcal{R})) \leq \mathcal{E}_{\lambda_2}(\pi(\mathcal{R}))$. Subtracting on each side, it comes
$$\mathcal{E}_{\lambda_1}(\pi(\mathcal{R})) - \mathcal{E}_{\lambda_1}(\mathcal{R}) \leq \mathcal{E}_{\lambda_2}(\pi(\mathcal{R})) - \mathcal{E}_{\lambda_2}(\mathcal{R})\ . \tag{A.3}$$
Finally, $\mathcal{E}_{\lambda_1}(\pi(\mathcal{R})) - \mathcal{E}_{\lambda_1}(\mathcal{R}) \geq 0$ implies $\mathcal{E}_{\lambda_2}(\pi(\mathcal{R})) - \mathcal{E}_{\lambda_2}(\mathcal{R}) \geq 0$, or, put differently,
$$\mathcal{E}_{\lambda_1}(\pi(\mathcal{R})) \geq \mathcal{E}_{\lambda_1}(\mathcal{R}) \Rightarrow \mathcal{E}_{\lambda_2}(\pi(\mathcal{R})) \geq \mathcal{E}_{\lambda_2}(\mathcal{R}) \tag{A.4}$$
which is the definition of scale-increasingness [101, 103] for the energy function $\mathcal{E}_\lambda$.
3. **Ordering of the optimal cuts $\{\pi^\star_\lambda\}$.** Consider finally $0 \leq \lambda_1 \leq \lambda_2$. There exist two optimal cuts $\pi^\star_{\lambda_1}$ and $\pi^\star_{\lambda_2}$ which can be obtained by conducting Bellman's dynamic program over the hierarchy H. Take $\mathcal{R} \in \pi^\star_{\lambda_1}$, $\mathcal{R}$ has a lower energy than any of its partial partitions $\pi(\mathcal{R}) \in \Pi_E(H(\mathcal{R}))$, thus $\mathcal{E}_{\lambda_1}(\mathcal{R}) \leq \mathcal{E}_{\lambda_1}(\pi(\mathcal{R}))$. Therefore, $\mathcal{E}_{\lambda_2}(\mathcal{R}) \leq \mathcal{E}_{\lambda_2}(\pi(\mathcal{R}))$ by scale-increasingness of $\mathcal{E}_\lambda$, meaning that $\mathcal{R}$ is temporary optimal for $\mathcal{E}_{\lambda_2}$, and thus either $\mathcal{R} \in \pi^\star_{\lambda_2}$ or $\mathcal{R} \subset \mathcal{R}' \in \pi^\star_{\lambda_2}$. As this holds for any $\mathcal{R} \in \pi^\star_{\lambda_1}$, one finally has $\pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2}$, which achieves the proof.

□

B Derivation of the Generalized Likelihood Ratio Test for hyperspectral change detection

Let $\mathbf{x} \in \mathbb{R}^N$ be some $N-$dimensional vector (corresponding to the difference of two hyperpsectral frames in the context of chapter 3), such that the two competing hypotheses for the distribution of $\mathbf{x}$ are a multivariate Gaussian distribution with covariance matrix Σ, either with a zero mean $\mathbf{0}$ (being the null hypothesis $\mathbf{H}_0$) or an unknown mean $\boldsymbol{\mu} \neq \mathbf{0}$ (the alternative hypothesis $\mathbf{H}_1$:

$$\begin{aligned} &\mathbf{H}_0 : f(\mathbf{x}|\mathbf{H}_0) \sim \mathcal{N}(\mathbf{0}, \Sigma) \\ &\mathbf{H}_1 : f(\mathbf{x}|\mathbf{H}_1) \sim \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right), \text{ with } \boldsymbol{\mu} \neq \mathbf{0} \text{ unknown.} \end{aligned} \tag{B.5}$$

Testing whether $\mathbf{x}$ is more likely to follow $\mathbf{H}_0$ or $\mathbf{H}_1$ classically involves a Likelihood Ratio Test, known according to the Neyman-Pearson lemma, to be the most powerful test for a given probability of false alarm. As $\boldsymbol{\mu}$ is unknown in the alternative hypothesis $\mathbf{H}_1$ however, the test (B.5) is solved using the Generalized Likelihood Ratio Test (GLRT). Assuming that we have S samples $\mathbf{x}_i, i = 1, \ldots, S$ (being the neighbors of $\mathbf{x}$), the GLRT write:

$$\Lambda(\mathbf{x}) = \frac{\max\limits_{\boldsymbol{\mu}\neq 0} \prod\limits_{i=1}^{S} \left[f(\mathbf{x}_i|\mathbf{H}_1)\right]}{\prod\limits_{i=1}^{S} \left[f(\mathbf{x}_i|\mathbf{H}_0)\right]} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \ . \tag{B.6}$$

It is known that the numerator of (B.6) is maximized when the unknown $\boldsymbol{\mu}$ is taken to be the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\mu}}$ of the actual mean value, given by:

$$\hat{\boldsymbol{\mu}} = \frac{1}{S} \sum_{i=1}^{S} \mathbf{x}_i \ . \tag{B.7}$$

Replacing $f(\mathbf{x}|\mathbf{H}_0)$ and $f(\mathbf{x}|\mathbf{H}_1)$ in (B.6) by their analytical expression gives:

$$\Lambda(\mathbf{x}) = \frac{\prod\limits_{i=1}^{S} \left[\frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right)\right]}{\prod\limits_{i=1}^{S} \left[\frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i\right)\right]} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \tag{B.8}$$

where $|\Sigma|$ is the determinant of the covariance matrix Σ. After some simplifications, one gets

$$\Lambda(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}\sum\limits_{i=1}^{S}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right]}{\exp\left[-\frac{1}{2}\sum\limits_{i=1}^{S}\mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i\right]} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}} \tag{B.9}$$

which rewrites as

$$\Lambda(\mathbf{x}) = \exp\left[-\frac{1}{2}\sum_{i=1}^{S}\left[(\mathbf{x}_i-\boldsymbol{\mu})^T\,\Sigma^{-1}\,(\mathbf{x}_i-\boldsymbol{\mu}) - \mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\right]\right] \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}}\ . \tag{B.10}$$

Developing the inner bracket gives

$$\Lambda(\mathbf{x}) = \exp\left[-\frac{1}{2}\sum_{i=1}^{S}\left[\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\Sigma^{-1}\mathbf{x}_i\right]\right] \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}}\ . \tag{B.11}$$

Taking the log on both sides yields

$$\Lambda(\mathbf{x}) = \sum_{i=1}^{S}\boldsymbol{\mu}^T\Sigma^{-1}\mathbf{x}_i - \frac{S}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}}\,[2]. \tag{B.12}$$

As the actual value of $\boldsymbol{\mu}$ does not matter, it is possible here to move the term $\frac{S}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}$ to the right side of the equation to get

$$\Lambda(\mathbf{x}) = \boldsymbol{\mu}^T\Sigma^{-1}\left(\sum_{i=1}^{S}\mathbf{x}_i\right) \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}}\ . \tag{B.13}$$

At this point however, the test is not computable in practice as $\boldsymbol{\mu}$ is unknown. Replacing $\boldsymbol{\mu}$ by its MLE $\hat{\boldsymbol{\mu}} = \frac{1}{S}\sum_{i=1}^{S}\mathbf{x}_i$ according to equation (B.7), or alternatively, replacing $\sum_{i=1}^{S}\mathbf{x}_i$ by $S\hat{\boldsymbol{\mu}}$ in (B.13) eventually yields to the final expression of the GLRT

$$\Lambda(\mathbf{x}) = S\hat{\boldsymbol{\mu}}^T\Sigma^{-1}\hat{\boldsymbol{\mu}} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \gamma_{\text{GLRT}}\ . \tag{B.14}$$

2. With a slight abuse of notation, the threshold is still denoted γ_{GLRT}.

C Composing a braid using cuts of independent hierarchies

Let $B = \{\pi_1^1, \pi_1^2, \pi_2^1, \pi_2^2\}$ be a family of partitions composed following the procedure described in section 4.4.1. $\pi_{i,j}^{k,l} = \pi_i^k \vee \pi_j^l$ denotes the pairwise refinement suprema of B. In particular, the 4 partitions composing B generates $\binom{4}{2} = 6$ different pairwise refinement suprema $\pi_{1,1}^{1,2}, \pi_{1,2}^{1,1}, \pi_{1,2}^{1,2}, \pi_{1,2}^{2,1}, \pi_{1,2}^{2,2}, \pi_{2,2}^{1,2}$. Checking that B is a braid amounts to verify whether the $\pi_{i,j}^{k,l}$ all defines cuts of the same monitor hierarchy H_m, which is equivalent to showing that they are (at least) all h-equivalent to each other. In order to show the braid structure of B, we first demonstrate the following result:

Lemma 1
Let $\pi_1, \pi_2, \pi_3 \in \Pi_E$ be some partitions of E such that $\pi_1 \overset{h}{\simeq} \pi_3$ and $\pi_2 \leq \pi_3$. Then $\pi_1 \vee \pi_2 \overset{h}{\simeq} \pi_3$.

Proof. In the most general case where π_1 and π_3 are h-equivalent but can nonetheless not be ordered, it means that π_1 is a refinement of π_3 in some parts of E, and is refined by π_3 in the other parts. In the former case, let $\mathcal{R}_3$ be a region of π_3 and $\pi_1(\mathcal{R}_3), \pi_2(\mathcal{R}_3)$ be the refinements (partial partitions) of $\mathcal{R}_3$ in π_1 and π_2. Then, $\pi_1(\mathcal{R}_3) \vee \pi_2(\mathcal{R}_3)$ is also a refinement of $\mathcal{R}_3$, implying that $\pi_1 \vee \pi_2$ refines π_3 in the part of E covered by $\mathcal{R}_3$. In the case where π_3 is locally a refinement of π_1, then given $\mathcal{R}_1 \in \pi_1$, there exists a refinement $\pi_3(\mathcal{R}_1)$ of $\mathcal{R}_1$ in π_3, and therefore a refinement $\pi_2(\mathcal{R}_1)$ of $\mathcal{R}_1$ in π_2 since $\pi_2 \leq \pi_3$. Therefore, $\{\mathcal{R}_1\} \vee \pi_2(\mathcal{R}_1) = \{\mathcal{R}_1\}$ and thus π_3 refines $\pi_1 \vee \pi_2$ in the part of E covered by $\mathcal{R}_1$. Finally, $\pi_1 \vee \pi_2$ either refines or is refined by π_3 in all parts of E, hence $\pi_1 \vee \pi_2 \overset{h}{\simeq} \pi_3$. □

Following, we prove that the pairwise refinement suprema of B are organized as displayed by figure 4.7b, which, combined with the above lemma 1, demonstrates the braid structure of B.

- $\pi_{1,2}^{1,2} = \pi_1^1 \vee \pi_2^2 = \pi_1^1$ by construction of B. Similarly, $\pi_{1,2}^{2,1} = \pi_2^1$. By definition of the refinement suprema, they are both refinements of $\pi_{1,2}^{1,1} = \pi_1^1 \vee \pi_2^1$.
- $\pi_{1,2}^{1,2} = \pi_1^1 \overset{h}{\simeq} \pi_2^1 = \pi_{1,2}^{2,1}$ by construction of B.
- $\pi_1^1 \overset{h}{\simeq} \pi_1^2$ as they are both cuts of the same hierarchy H_1, so is their supremum $\pi_{1,1}^{1,2}$. This implies that $\pi_{1,1}^{1,2} \overset{h}{\simeq} \pi_{1,2}^{1,2}$. With the same argument, one has $\pi_{2,2}^{1,2} \overset{h}{\simeq} \pi_{1,2}^{2,1}$.
- $\pi_{1,2}^{2,2} = \pi_1^2 \vee \pi_2^2$, with $\pi_1^2 \overset{h}{\simeq} \pi_1^1$ and $\pi_2^2 \leq \pi_1^1$. Using the lemma 1, it follows that $\pi_{1,2}^{2,2} \overset{h}{\simeq} \pi_1^1 = \pi_{1,2}^{1,2}$. With the same argument, one proves that $\pi_{1,2}^{2,2} \overset{h}{\simeq} \pi_{1,2}^{2,1}$.
- Again, $\pi_{1,2}^{2,2} = \pi_1^2 \vee \pi_2^2$ and $\pi_2^2 \leq \pi_1^1$ by construction of B. Therefore $\pi_{1,2}^{2,2} \leq \pi_1^1 \vee \pi_1^2 = \pi_{1,1}^{1,2}$. With the same argument, one shows that $\pi_{1,2}^{2,2} \leq \pi_{2,2}^{1,2}$.
- Finally, $\pi_1^1 \overset{h}{\simeq} \pi_2^1$ and $\pi_1^1 \geq \pi_2^2$ by construction of B. Using the lemma 1, it follows that $\pi_1^1 \overset{h}{\simeq} \pi_{2,2}^{1,2}$. In addition, $\pi_1^2 \leq \pi_2^1$, thus $\pi_1^2 \leq \pi_2^1 \vee \pi_2^2 = \pi_{2,2}^{1,2}$. Using the lemma 1 again for π_1^1, π_1^2 and $\pi_{2,2}^{1,2}$, it follows that $\pi_{1,1}^{1,2} \overset{h}{\simeq} \pi_{2,2}^{1,2}$.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[2] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[3] H. G. Akcay and S. Aksoy. Automatic detection of geospatial objects using multiple hierarchical segmentations. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(7):2097–2111, 2008.

[4] I. F. Akyildiz, D. Pompili, and T. Melodia. Underwater acoustic sensor networks: research challenges. *Ad hoc networks*, 3(3):257–279, 2005.

[5] A. Alonso-González, C. López-Martínez, and P. Salembier. Filtering and segmentation of polarimetric SAR data based on binary partition trees. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(2):593–605, 2012.

[6] A. Alonso-González, S. Valero, J. Chanussot, C. López-Martínez, and P. Salembier. Processing multidimensional SAR and hyperspectral images with binary partition tree. *Proceedings of the IEEE*, 101(3):723–747, 2013.

[7] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini. A global quality measurement of pan-sharpened multispectral imagery. *Geoscience and Remote Sensing Letters, IEEE*, 1(4):313–317, 2004.

[8] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce. Comparison of pansharpening algorithms: Outcome of the 2006 GRSS data-fusion contest. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(10):3012–3021, 2007.

[9] T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 2000.

[10] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011.

[11] G. P. Asner, D. E. Knapp, T. Kennedy-Bowdoin, M. O. Jones, R. E. Martin, J. Boardman, and C. B. Field. Carnegie airborne observatory: in-flight fusion of hyperspectral imaging and waveform light detection and ranging for three-dimensional studies of ecosystems. *Journal of Applied Remote Sensing*, 1(1):013536–013536, 2007.

[12] M. Baatz and A. Schäpe. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In *Angewandte Geographische Informationsverarbeitung XII*, pages 12–23. Wichmann-Verlag, Heidelberg, 2000.

[13] A. Baldridge, S. Hook, C. Grove, and G. Rivera. The ASTER spectral library version 2.0. *Remote Sensing of Environment*, 113(4):711–715, 2009.

[14] C. Ballester, V. Caselles, L. Igual, and L. Garrido. Level lines selection with variational models for segmentation and encoding. *Journal of Mathematical Imaging and Vision*, 27(1):5–27, 2007.

[15] A. Banerjee, P. Burlina, and J. Broadwater. Hyperspectral video for illumination-invariant tracking. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS'09. First Workshop on*, pages 1–4. IEEE, 2009.

[16] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, volume 5, 2006.

[17] Y. Bazi, L. Bruzzone, and F. Melgani. An unsupervised approach based on the generalized gaussian model to automatic change detection in multitemporal SAR images. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(4):874–887, 2005.

[18] J.-M. Beaulieu and M. Goldberg. Hierarchy in picture segmentation: A stepwise optimization approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(2):150–163, 1989.

[19] J. A. Benediktsson, M. Pesaresi, and K. Amason. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(9):1940–1949, 2003.

[20] C. F. Bennström and J. R. Casas. Binary-partition-tree creation using a quasi-inclusion criterion. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 259–264. IEEE, 2004.

[21] C. Berger, M. Voltersen, R. Eckardt, J. Eberle, T. Heyer, N. Salepci, S. Hese, C. Schmullius, J. Tao, S. Auer, et al. Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS data fusion contest. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(3):1324–1340, 2013.

[22] J. Bioucas-Dias and J. Nascimento. Hyperspectral Subspace Identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445, 2008.

[23] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(2):354–379, 2012.

[24] J. L. Bishop, E. Z. N. Dobrea, N. K. McKeown, M. Parente, B. L. Ehlmann, J. R. Michalski, R. E. Milliken, F. Poulet, G. A. Swayze, J. F. Mustard, et al. Phyllosilicate diversity and past aqueous activity revealed at Mawrth Vallis, Mars. *Science*, 321(5890):830–833, 2008.

[25] J. Black, T. Ellis, and P. Rosin. A novel method for video tracking performance evaluation. In *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 125–132, 2003.

[26] F. Bovolo and L. Bruzzone. A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(6):1658–1670, 2007.

[27] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[28] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

[29] J. B. Broadwater, D. Limsui, and A. K. Carr. A Primer for Chemical Plume Detection Using LWIR Sensors. Technical report, John Hopkins Applied Physics Laboratory, 04 2011.

[30] J. B. Broadwater, T. S. Spisz, and A. K. Carr. Detection of gas plumes in cluttered environments using long-wave infrared hyperspectral sensors. In *SPIE Defense and Security Symposium*, pages 69540R–69540R. International Society for Optics and Photonics, 2008.

[31] L. Bruzzone and D. F. Prieto. Automatic analysis of the difference image for unsupervised change detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 38(3):1171–1182, 2000.

[32] L. Bruzzone and D. F. Prieto. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(2):456–460, 2001.

[33] L. Bruzzone, D. F. Prieto, and S. B. Serpico. A neural-statistical approach to multitemporal and multisource remote-sensing image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(3):1350–1359, 1999.

[34] A. Buades, B. Coll, and J.-M. Morel. Nonlocal image and movie denoising. *International journal of computer vision*, 76(2):123–139, 2008.

[35] P. Bunting and R. Lucas. The delineation of tree crowns in australian mixed species forests using hyperspectral compact airborne spectrographic imager (CASI) data. *Remote Sensing of Environment*, 101(2):230–248, 2006.

[36] F. Calderero and F. Marques. Region merging techniques using information theory statistical measures. *Image Processing, IEEE Transactions on*, 19(6):1567–1586, 2010.

[37] V. D. Calhoun and T. Adali. Feature-based fusion of medical imaging data. *Information Technology in Biomedicine, IEEE Transactions on*, 13(5):711–720, 2009.

[38] M. A. Calin, S. V. Parasca, D. Savastru, and D. Manea. Hyperspectral imaging in the medical field: present and future. *Applied Spectroscopy Reviews*, 49(6):435–447, 2014.

[39] G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(6):1351–1362, 2005.

[40] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, and M. Martínez-Ramón. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(6):1822–1835, 2008.

[41] J. Cardelino, G. Randall, M. Bertalmio, and V. Caselles. Region based segmentation using the tree of shapes. In *Image Processing, 2006 IEEE International Conference on*, pages 2421–2424. IEEE, 2006.

[42] E. Carlinet and T. Géraud. A comparative review of component tree computation algorithms. *Image Processing, IEEE Transactions on*, 23(9):3885–3895, 2014.

[43] E. Carlinet and T. Géraud. A morphological tree of shapes for color images. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1132–1137. IEEE, 2014.

[44] E. Carlinet and T. Géraud. A color tree of shapes with illustrations on filtering, simplification, and segmentation. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 363–374. Springer, 2015.

[45] V. Caselles, B. Coll, and J.-M. Morel. Topographic maps and local contrast changes in natural images. *International Journal of Computer Vision*, 33(1):5–27, 1999.

[46] A. Chambolle. Finite-differences discretizations of the Mumford-Shah functional. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(02):261–288, 1999.

[47] T. E. Chan, L. Vese, et al. A level set algorithm for minimizing the Mumford-Shah functional in image processing. In *Variational and Level Set Methods in Computer Vision, 2001. Proceedings. IEEE Workshop on*, pages 161–168. IEEE, 2001.

[48] C.-M. Chen, G. Hepner, and R. Forster. Fusion of hyperspectral and radar data using the IHS transformation to enhance urban surface features. *ISPRS Journal of photogrammetry and Remote Sensing*, 58(1):19–30, 2003.

[49] J. M. Chen, J. Liu, S. G. Leblanc, R. Lacaze, and J.-L. Roujean. Multi-angular optical remote sensing for assessing vegetation structure and carbon absorption. *Remote Sensing of Environment*, 84(4):516–525, 2003.

[50] T. Chen. Audiovisual speech processing. *Signal Processing Magazine, IEEE*, 18(1):9–21, 2001.

[51] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2):713–718, 1992.

[52] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[53] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[54] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[55] T. F. Cox and M. A. Cox. *Multidimensional scaling*. CRC Press, 2000.

[56] N. Cvejic, D. Bull, and N. Canagarajah. Region-based multimodal image fusion using ICA bases. *Sensors Journal, IEEE*, 7(5):743–751, 2007.

[57] K. Dabov, A. Foi, and K. Egiazarian. Video denoising by sparse 3D transform-domain collaborative filtering. In *Proc. 15th European Signal Processing Conference*, volume 1, pages 145–149, 2007.

[58] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[59] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, and J. Chanussot. Challenges and opportunities of multimodality and Data Fusion in Remote Sensing. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European*, pages 106–110. IEEE, 2014.

[60] M. Dalponte, L. Bruzzone, and D. Gianelle. Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(5):1416–1427, 2008.

[61] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici. Hyperspectral and lidar data fusion: Outcome of the 2013 GRSS data fusion contest. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7:2405–2418, 2014.

[62] J. D'Errico. Interpolate (& extrapolates) nan elements in a 2d array., 2004. (http://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint-nans), MATLAB Central File Exchange. Retrieved August 13, 2012.

[63] D. J. Diner, J. C. Beckert, T. H. Reilly, C. J. Bruegge, J. E. Conel, R. A. Kahn, J. V. Martonchik, T. P. Ackerman, R. Davies, S. A. Gerstl, et al. Multi-angle imaging spectroradiometer (MISR) instrument description and experiment overview. *Geoscience and Remote Sensing, IEEE Transactions on*, 36(4):1072–1087, 1998.

[64] D. L. Donoho. CART and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870–1911, 1997.

[65] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5):352–359, 2002.

[66] L. Drumetz, M. A. Veganzones, R. Marrero, G. Tochon, M. Dalla Mura, A. Plaza, and J. Chanussot. Binary partition tree-based local spectral unmixing. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, pages n–c, 2014.

[67] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on*, 2(3):141–151, 2000.

[68] O. Eches, N. Dobigeon, and J.-Y. Tourneret. Enhancing hyperspectral image unmixing with spatial correlations. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(11):4239–4247, 2011.

[69] G. Elmasry, M. Kamruzzaman, D.-W. Sun, and P. Allen. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. *Critical Reviews in Food Science and Nutrition*, 52(11):999–1023, 2012.

[70] V. Farley, A. Vallières, M. Chamberland, A. Villemaire, and J.-F. Legault. Performance of the FIRST: a long-wave infrared hyperspectral imaging sensor. In *Optics/Photonics in Security and Defence*, pages 63980T–63980T. International Society for Optics and Photonics, 2006.

[71] V. Farley, A. Vallières, A. Villemaire, M. Chamberland, P. Lagueux, and J. Giroux. Chemical agent detection and identification with a hyperspectral imaging infrared sensor.

In *Optics/Photonics in Security and Defence*, pages 673918–673918. International Society for Optics and Photonics, 2007.

[72] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(11):3804–3814, 2008.

[73] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, 2013.

[74] J.-B. Féret and G. P. Asner. Tree species discrimination in tropical forests using airborne imaging spectroscopy. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):73–84, 2013.

[75] R. A. Finkel and J. L. Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974.

[76] L. Franek, D. D. Abdala, S. Vega-Pons, and X. Jiang. Image segmentation fusion using general ensemble clustering methods. In *Computer Vision–ACCV 2010*, pages 373–384. Springer, 2011.

[77] S. Gao and T. D. Bui. Image segmentation and selective smoothing by using Mumford-Shah model. *Image Processing, IEEE Transactions on*, 14(10):1537–1549, 2005.

[78] L. Garrido, P. Salembier, and D. Garcia. Extensive operators in partition lattices for image sequence analysis. *Signal Processing*, 66(2):157–180, 1998.

[79] T. Gerhart, J. Sunu, L. Lieu, E. Merkurjev, J.-M. Chang, J. Gilles, and A. L. Bertozzi. Detection and tracking of gas plumes in LWIR hyperspectral video sequence data. In *SPIE Defense, Security, and Sensing*, pages 87430J–87430J. International Society for Optics and Photonics, 2013.

[80] A. Gillespie, S. Rokugawa, T. Matsunaga, J. S. Cothern, S. Hook, and A. B. Kahle. A temperature and emissivity separation algorithm for Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images. *Geoscience and Remote Sensing, IEEE Transactions on*, 36(4):1113–1126, 1998.

[81] R. Goecke. Current trends in joint audio-video signal processing: a review. In *ISSPA*, pages 70–73, 2005.

[82] A. F. Goetz. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote Sensing of Environment*, 113:S5–S16, 2009.

[83] M. Govender, K. Chetty, and H. Bulcock. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa*, 33(2):145 – 152, 2007.

[84] M. Graña and M. Veganzones. An endmember-based distance for content based hyperspectral image retrieval. *Pattern Recognition*, 45(9):3472–3489, 2012.

[85] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, et al. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 65(3):227–248, 1998.

[86] L. Guigues. *Modèles multi-échelles pour la segmentation d'images*. PhD thesis, Université de Cergy-Pontoise, 2003.

[87] L. Guigues, J. P. Cocquerez, and H. Le Men. Scale-sets image analysis. *International Journal of Computer Vision*, 68(3):289–317, 2006.

[88] V. C. Gungor, B. Lu, and G. P. Hancke. Opportunities and challenges of wireless sensor networks in smart grid. *Industrial Electronics, IEEE Transactions on*, 57(10):3557–3564, 2010.

[89] D. Heinz and C.-I. Chang. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):529–545, 2001.

[90] M. Hilbert and P. López. The world's technological capacity to store, communicate, and compute information. *science*, 332(6025):60–65, 2011.

[91] E. Hirsch and E. Agassi. Detection of gaseous plumes in IR hyperspectral images using hierarchical clustering. *Applied optics*, 46(25):6368–6374, 2007.

[92] R. Horaud and O. Monga. *Vision par ordinateur: outils fondamentaux*. Editions Hermès, 1995.

[93] H. Hu, J. Sunu, and A. L. Bertozzi. Multi-class graph mumford-shah model for plume detection using the MBO scheme. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 209–222. Springer, 2015.

[94] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza. Total variation spatial regularization for sparse hyperspectral unmixing. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(11):4484–4502, 2012.

[95] S. Jia and Y. Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(1):161–173, 2009.

[96] R. Jones. Component trees for image filtering and segmentation. In *Proceedings of the 1997 IEEE Workshop on Nonlinear Signal and Image Processing, Mackinac Island*, 1997.

[97] R. Jones. Connected filtering and segmentation using component trees. *Computer Vision and Image Understanding*, 75(3):215–228, 1999.

[98] N. Keshava and J. Mustard. Spectral unmixing. *Signal Processing Magazine, IEEE*, 19(1):44–57, 2002.

[99] R. L. Kettig and D. Landgrebe. Classification of multispectral image data by extraction and classification of homogeneous objects. *Geoscience Electronics, IEEE Transactions on*, 14(1):19–26, 1976.

[100] N. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and computational harmonic analysis*, 10(3):234–253, 2001.

[101] B. R. Kiran. *Energetic-Lattice based optimization*. PhD thesis, Université Paris-Est, 2014.

[102] B. R. Kiran and J. Serra. Ground truth energies for hierarchies of segmentations. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 123–134. Springer, 2013.

[103] B. R. Kiran and J. Serra. Global–local optimizations by hierarchical cuts and climbing energies. *Pattern Recognition*, 47(1):12–24, 2014.

[104] B. R. Kiran and J. Serra. Braids of partitions. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 217–228. Springer, 2015.

[105] J. D. Kushla and W. J. Ripple. Assessing wildfire effects with landsat thematic mapper data. *International Journal of Remote Sensing*, 19(13):2493–2507, 1998.

[106] D. Lahat, T. Adali, and C. Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[107] D. Lahat, T. Adaly, and C. Jutten. Challenges in multimodal data fusion. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European*, pages 101–105. IEEE, 2014.

[108] S. Lakshmanan and H. Derin. Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(8):799–813, 1989.

[109] E. R. Larrieux. Performance evaluation of chemical plume detection and quantification algorithms. Master's thesis, Northeastern University, 2009.

[110] C. Lawson. *Solving Least Squares Problems*. Prentice Hall, 1974.

[111] D. Le Gall. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991.

[112] J. Li, J. Wang, and G. Wiederhold. IRM: Integrated Region Matching for Image Retrieval. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, pages 147–156, New York, NY, USA, 2000. ACM.

[113] S. Z. Li. *Markov random field modeling in image analysis*, volume 26. Springer, 2009.

[114] W. Liao, X. Huang, F. Van Coillie, S. Gautama, A. Pizurica, W. Philips, H. Liu, T. Zhu, M. Shimoni, G. Moser, and D. Tuia. Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS datafusion contest. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, (submitted), 2015.

[115] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, C. Thiel, J. Inglada, E. Christophe, J. Chanussot, et al. Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRSS data fusion contest. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(11):3857–3865, 2009.

[116] G. A. Licciardi, A. Villa, M. Dalla Mura, L. Bruzzone, J. Chanussot, and J. A. Benediktsson. Retrieval of the height of buildings from WorldView-2 multi-angular imagery using attribute filters and geometric invariant moments. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(1):71–79, 2012.

[117] K. Lim, P. Treitz, M. Wulder, B. St-Onge, and M. Flood. LiDAR remote sensing of forest structure. *Progress in physical geography*, 27(1):88–106, 2003.

[118] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 61, 2009.

[119] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 246–253. IEEE, 2006.

[120] Z. Liu, L. Shen, and Z. Zhang. Unsupervised image segmentation based on analysis of binary partition tree for salient object extraction. *Signal Processing*, 91(2):290–299, 2011.

[121] S. P. Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[122] L. Loncan, J. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. Licciardi, M. Simoes, J. Tourneret, et al. Introducing hyperspectral pansharpening. *IEEE Geoscience and Remote Sensing Magazine (submitted)*, 2015.

[123] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, et al. Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009–2010 data fusion contest. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(1):331–342, 2012.

[124] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco. Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food and Bioprocess Technology*, 5(4):1121–1142, 2012.

[125] G. Lu and B. Fei. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901–010901, 2014.

[126] H. Lu, J. C. Woods, and M. Ghanbari. Binary partition tree analysis based on region evolution and its application to tree simplification. *Image Processing, IEEE Transactions on*, 16(4):1131–1138, 2007.

[127] H. Lu, J. C. Woods, and M. Ghanbari. Binary partition tree for semantic object extraction and image segmentation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3):378–383, 2007.

[128] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *Medical Imaging, IEEE Transactions on*, 16(2):187–198, 1997.

[129] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference (BMVC)*, 2007.

[130] D. Manolakis, S. Golowich, and R. DiPietro. Long-Wave Infrared Hyperspectral Remote Sensing of Chemical Clouds: A focus on signal processing approaches. *Signal Processing Magazine, IEEE*, 31(4):120–141, 2014.

[131] D. Manolakis and G. Shaw. Detection algorithms for hyperspectral imaging applications. *Signal Processing Magazine, IEEE*, 19(1):29–43, 2002.

[132] D. G. Manolakis and F. M. D'Amico. A taxonomy of algorithms for chemical vapor detection with hyperspectral imaging spectroscopy. In *Defense and Security*, pages 125–133. International Society for Optics and Photonics, 2005.

[133] D. G. Manolakis, G. A. Shaw, and N. Keshava. Comparative analysis of hyperspectral adaptive matched filter detectors. In *AeroSense 2000*, pages 2–17. International Society for Optics and Photonics, 2000.

[134] G. Martín and A. Plaza. Region-based spatial preprocessing for endmember extraction and spectral unmixing. *Geoscience and Remote Sensing Letters, IEEE*, 8(4):745–749, 2011.

[135] G. Martin and A. Plaza. Spatial-spectral preprocessing prior to endmember identification and unmixing of remotely sensed hyperspectral data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(2):380–395, 2012.

[136] P. Mather and M. Koch. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Sons, 2011.

[137] E. Merkurjev, T. Kostic, and A. L. Bertozzi. An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.

[138] E. Merkurjev, J. Sunu, and A. L. Bertozzi. Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 689–693. IEEE, 2014.

[139] F. Meyer and P. Maragos. Nonlinear scale-space representation with morphological levelings. *Journal of Visual Communication and Image Representation*, 11(2):245–265, 2000.

[140] P. Monasse and F. Guichard. Fast computation of a contrast-invariant image representation. *IEEE Transactions on Image Processing*, 9(5):860–872, 2000.

[141] R. J. Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[142] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.

[143] L. Najman, J. Cousty, and B. Perret. Playing with Kruskal: algorithms for morphological trees in edge-weighted graphs. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 135–146. Springer, 2013.

[144] J. Nascimento and J. Dias. Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

[145] C. J. Needham and R. D. Boyle. *Performance evaluation metrics and statistics for positional tracker evaluation*. Springer, 2003.

[146] G. Noyel, J. Angulo, and D. Jeulin. Morphological segmentation of hyperspectral images. *Image Anal. Stereol*, 26(3):101–109, 2007.

[147] E. M. O'Donnell, D. W. Messinger, C. Salvaggio, and J. R. Schott. Identification and detection of gaseous effluents from hyperspectral imagery using invariant algorithms. In *Defense and Security*, pages 573–582. International Society for Optics and Photonics, 2004.

[148] G. K. Ouzounis and P. Soille. The alpha-tree algorithm. Technical report, JRC Scientific and Policy Report, European Commission, Joint Research Centre, 2012.

[149] Pacific Northwest National Laboratory IR Database. http://nwir.pnl.gov, April 2014.

[150] F. Pacifici, J. Chanussot, and Q. Du. 2011 GRSS data fusion contest: Exploiting WorldView-2 multi-angular acquisitions. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, pages 1163–1166. IEEE, 2011.

[151] F. Pacifici, F. Del Frate, W. J. Emery, P. Gamba, and J. Chanussot. Urban mapping using coarse SAR and optical data: Outcome of the 2007 GRSS data fusion contest. *Geoscience and Remote Sensing Letters, IEEE*, 5(3):331–335, 2008.

[152] P. Padhy, K. Martinez, A. Riddoch, H. Ong, and J. K. Hart. Glacial environment monitoring using sensor networks. In *Proceedings of the First Real-World Wireless Sensor Networks Workshop (REALWSN'05)*, pages 10–14, 2005.

[153] G. Palou and P. Salembier. Hierarchical video representation with trajectory binary partition tree. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2099–2106. IEEE, 2013.

[154] Y. Pan, J. D. Birdwell, and S. M. Djouadi. Preferential image segmentation using trees of shapes. *Image Processing, IEEE Transactions on*, 18(4):854–866, 2009.

[155] A. Pardo. Semantic image segmentation using morphological tools. In *Image Processing (ICIP), IEEE International Conference on*, pages 745–748, 2002.

[156] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information fusion*, 4(4):259–280, 2003.

[157] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, et al. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:S110–S122, 2009.

[158] A. Plaza, P. Martinez, R. Pérez, and J. Plaza. Spatial–spectral endmember extraction by multidimensional morphological operations. *Geoscience and Remote Sensing, IEEE Transactions on*, 40(9):2025–2041, 2002.

[159] G. Priestnall, J. Jaafar, and A. Duncan. Extracting urban features from LiDAR digital surface models. *Computers, Environment and Urban Systems*, 24(2):65–78, 2000.

[160] R. Pu, P. Gong, Y. Tian, X. Miao, R. I. Carruthers, and G. L. Anderson. Invasive species change detection using artificial neural networks and CASI hyperspectral imagery. *Environmental monitoring and assessment*, 140(1-3):15–32, 2008.

[161] Z. Qi, A. G.-O. Yeh, X. Li, and Z. Lin. A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data. *Remote Sensing of Environment*, 118:21–39, 2012.

[162] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly. Hyperspectral unmixing via sparsity-constrained nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(11):4282–4297, 2011.

[163] J. Qin, T. Laurent, K. Bui, R. V. Tan, J. Dahilig, S. Wang, J. Rohe, J. Sunu, and A. L. Bertozzi. Detecting plumes in LWIR using robust nonnegative matrix factorization with graph-based initialization. In *SPIE Defense+ Security*, pages 94720V–94720V. International Society for Optics and Photonics, 2015.

[164] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *Image Processing, IEEE Transactions on*, 14(3):294–307, 2005.

[165] J. F. Randrianasoa, C. Kurtz, É. Desjardin, and N. Passat. Multi-image Segmentation: A Collaborative Approach Based on Binary Partition Trees. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 253–264. Springer, 2015.

[166] A. C. Rencher and W. F. Christensen. *Methods of multivariate analysis*, volume 709. John Wiley & Sons, 2012.

[167] J. Rogan, J. Franklin, and D. A. Roberts. A comparison of methods for monitoring multitemporal vegetation change using thematic mapper imagery. *Remote Sensing of Environment*, 80(1):143–156, 2002.

[168] D. Rogge, B. Rivard, J. Zhang, A. Sanchez, J. Harris, and J. Feng. Integration of spatial–spectral information for the improved extraction of endmembers. *Remote Sensing of Environment*, 110(3):287–303, 2007.

[169] T. Rohlfing and C. R. Maurer Jr. Shape-based averaging. *Image Processing, IEEE Transactions on*, 16(1):153–161, 2007.

[170] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[171] B. C. Russell, W. T. Freeman, A. Efros, J. Sivic, A. Zisserman, et al. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1605–1614. IEEE, 2006.

[172] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *Image Processing, IEEE Transactions on*, 9(4):561–576, 2000.

[173] P. Salembier, A. Oliveras, and L. Garrido. Antiextensive connected operators for image and sequence processing. *Image Processing, IEEE Transactions on*, 7(4):555–570, 1998.

[174] P. Salembier and J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *Image Processing, IEEE transactions on*, 4(8):1153–1160, 1995.

[175] O. Salerno, M. Pardàs, V. Vilaplana, and F. Marqués. Object recognition based on binary partition trees. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 2, pages 929–932. IEEE, 2004.

[176] L. L. Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991.

[177] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer, 2014.

[178] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 746–751. IEEE, 2000.

[179] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[180] G. Shalini, A. Pravin Renold, and B. Venkatalakshmi. Performance evaluation of path loss models for mobile Underwater Acoustic Sensor Networks. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*, pages 366–371. IEEE, 2012.

[181] S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick, and P. A. Johnson. Gas-phase databases for quantitative infrared spectroscopy. *Applied spectroscopy*, 58(12):1452–1461, 2004.

[182] E. Shusterman and M. Feder. Image compression via improved quadtree decomposition algorithms. *Image Processing, IEEE Transactions on*, 3(2):207–215, 1994.

[183] P. Soille. Constrained connectivity for hierarchical image partitioning and simplification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1132–1145, 2008.

[184] M. Spann and R. Wilson. A quad-tree approach to image segmentation which combines statistical and spatial information. *Pattern Recognition*, 18(3):257–269, 1985.

[185] T. S. Spisz, P. K. Murphy, C. C. Carter, A. K. Carr, A. Vallières, and M. Chamberland. Field test results of standoff chemical detection using the FIRST. In *Defense and Security Symposium*, pages 655408–655408. International Society for Optics and Photonics, 2007.

[186] D. W. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker. Anomaly detection from hyperspectral imagery. *Signal Processing Magazine, IEEE*, 19(1):58–69, 2002.

[187] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun. A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of neuroscience methods*, 204(1):68–81, 2012.

[188] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43(7):2367–2379, 2010.

[189] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson. SVM-and MRF-based method for accurate classification of hyperspectral images. *Geoscience and Remote Sensing Letters, IEEE*, 7(4):736–740, 2010.

[190] Y. Tarabalka, J. Tilton, J. Benediktsson, and J. Chanussot. A Marker-Based Approach for the Automated Selection of a Single Segmentation From a Hierarchical Set of Image Segmentations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):262–272, 2012.

[191] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(5):1301–1312, 2008.

[192] J. C. Tilton. A recursive PVM implementation of an image segmentation algorithm with performance results comparing the HIVE and the Cray T3E. In *Frontiers of Massively Parallel Computation, 1999. Frontiers' 99. The Seventh Symposium on the*, pages 146–153. IEEE, 1999.

[193] J. C. Tilton. Analysis of hierarchically related image segmentations. In *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on*, pages 60–69. IEEE, 2003.

[194] J. C. Tilton. Split-remerge method for eliminating processing window artifacts in recursive hierarchical segmentation, Apr. 13 2010. US Patent 7,697,759.

[195] G. Tochon, J. Chanussot, M. Dalla Mura, and A. L. Bertozzi. Hierarchical representation of hyperspectral video sequences: application to chemical gas plume tracking. *Pattern recognition*, (submitted).

[196] G. Tochon, J. Chanussot, J. Gilles, M. Dalla Mura, J.-M. Chang, and A. Bertozzi. Gas plume detection and tracking in hyperspectral video sequences using binary partition trees. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.

[197] G. Tochon, M. Dalla Mura, and J. Chanussot. Segmentation of Multimodal Images based on Hierarchies of Partitions. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 241–252. Springer, 2015.

[198] G. Tochon, M. Dalla Mura, M. A. Veganzones, and J. Chanussot. Braids of partition for the hierarchical analysis of multimodal images. *Pattern Recognition*, (submitted).

[199] G. Tochon, J. Féret, S. Valero, R. Martin, D. Knapp, P. Salembier, J. Chanussot, and G. Asner. On the use of binary partition trees for the tree crown segmentation of tropical rainforest hyperspectral images. *Remote Sensing of Environment*, 159:318–331, 2015.

[200] G. Tochon, J.-B. Feret, R. E. Martin, R. Tupayachi, J. Chanussot, and G. P. Asner. Binary partition tree as a hyperspectral segmentation tool for tropical rainforests. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 6368–6371. IEEE, 2012.

[201] D. M. Tralli, R. G. Blom, V. Zlotnicki, A. Donnellan, and D. L. Evans. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(4):185–198, 2005.

[202] E. Trucco and K. Plakas. Video tracking: a concise survey. *Oceanic Engineering, IEEE Journal of*, 31(2):520–529, 2006.

[203] A. Turlapaty, B. Gokaraju, Q. Du, N. H. Younan, and J. V. Aanstoos. A hybrid approach for building extraction from spaceborne multi-angular optical imagery. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(1):89–100, 2012.

[204] S. Valero. *Hyperspectral image processing and representation using Binary Partition Trees*. PhD thesis, Gipsa-Lab, Department of Images and Signals, Grenoble Institute of Technology, Grenoble, France, 2011.

[205] S. Valero, P. Salembier, and J. Chanussot. Comparison of merging orders and pruning strategies for binary partition tree in hyperspectral data. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2565–2568. IEEE, 2010.

[206] S. Valero, P. Salembier, and J. Chanussot. Hyperspectral image segmentation using binary partition trees. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1273–1276. IEEE, 2011.

[207] S. Valero, P. Salembier, and J. Chanussot. Hyperspectral image representation and processing with binary partition trees. *Image Processing, IEEE Transactions on*, 22(4):1430–1443, 2013.

[208] S. Valero, P. Salembier, and J. Chanussot. Object recognition in urban hyperspectral images using Binary Partition Tree representation. In *IGARSS*, pages 4098–4101, 2013.

[209] S. Valero, P. Salembier, and J. Chanussot. Object recognition in hyperspectral images using Binary Partition Tree representation. *Pattern Recognition Letters*, 56:45–51, 2015.

[210] S. Valero, P. Salembier, J. Chanussot, and C. M. Cuadras. Improved binary partition tree construction for hyperspectral images: application to object detection. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, pages 2515–2518. IEEE, 2011.

[211] A. Vallières, A. Villemaire, M. Chamberland, L. Belhumeur, V. Farley, J. Giroux, and J.-F. Legault. Algorithms for chemical detection, identification and quantification for thermal hyperspectral imagers. In *Optics East 2005*, pages 59950G–59950G. International Society for Optics and Photonics, 2005.

[212] F. D. Van der Meer, H. M. Van der Werff, F. J. van Ruitenbeek, C. A. Hecker, W. H. Bakker, M. F. Noomen, M. van der Meijde, E. J. M. Carranza, J. B. de Smeth, and T. Woldai. Multi-and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):112–128, 2012.

[213] H. Van Nguyen, A. Banerjee, and R. Chellappa. Tracking via object reflectance using a hyperspectral video camera. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 44–51. IEEE, 2010.

[214] M. Veganzones and M. Graña. Endmember Extraction Methods: A Short Review. In *Knowledge-Based Intelligent Information and Engineering Systems, 12th International Conference, KES 2008, Zagreb, Croatia, September 3-5, 2008, Proceedings, Part III*, volume 5179 of *Lecture Notes in Computer Science*, pages 400–407. Springer, 2008.

[215] M. Veganzones and M. Graña. A Spectral/Spatial CBIR System for Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):488–500, April 2012.

[216] M. A. Veganzones, G. Tochon, M. Dalla Mura, A. J. Plaza, and J. Chanussot. Hyperspectral image segmentation using a new spectral mixture-based binary partition tree representation. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 245–249. IEEE, 2013.

[217] M. A. Veganzones, G. Tochon, M. Dalla-Mura, A. J. Plaza, and J. Chanussot. Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation. *Image Processing, IEEE Transactions on*, 23(8):3574–3589, 2014.

[218] V. Vilaplana, F. Marques, and P. Salembier. Binary partition trees for object detection. *Image Processing, IEEE Transactions on*, 17(11):2201–2216, 2008.

[219] A. Villa, J. Chanussot, J. A. Benediktsson, and C. Jutten. Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution. *Selected Topics in Signal Processing, IEEE Journal of*, 5(3):521–533, 2011.

[220] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1(6):583–598, 1991.

[221] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[222] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[223] L. Wald. Some terms of reference in data fusion. *IEEE Transactions on geoscience and remote sensing*, 37(3):1190–1193, 1999.

[224] T. Wan, N. Canagarajah, and A. Achim. Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients. *Multimedia, IEEE Transactions on*, 11(4):624–633, 2009.

[225] P. Wattuya, K. Rothaus, J.-S. Praßni, and X. Jiang. A random walker based approach to combining multiple segmentations. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[226] G. Werner-Allen, J. Johnson, M. Ruiz, J. Lees, and M. Welsh. Monitoring volcanic eruptions with a wireless sensor network. In *Wireless Sensor Networks, 2005. Proceeedings of the Second European Workshop on*, pages 108–120. IEEE, 2005.

[227] M. Winter. N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Proceedings of SPIE*, volume 3753, pages 266–275, 1999.

[228] Y. Xu, T. Géraud, and L. Najman. Context-based energy estimator: Application to object segmentation on the tree of shapes. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1577–1580. IEEE, 2012.

[229] Y. Xu, T. Géraud, and L. Najman. Morphological filtering in shape spaces: Applications using tree-based image representations. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 485–488. IEEE, 2012.

[230] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13, 2006.

[231] F. Yin, D. Makris, and S. A. Velastin. Performance evaluation of object tracking algorithms. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil*, 2007.

[232] N. Yokoya, T. Yairi, and A. Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(2):528–537, 2012.

[233] S. Young. Detection and quantification of gases in industrial-stack plumes using thermal-infrared hyperspectral imaging. *Aerospace Report ATR-2002 (8407)*, 1, 2002.

[234] Y. Zhang. An alternating direction algorithm for nonnegative matrix factorization. Technical Report TR10-03, Rice University, 2010.

[235] S. Zhu and K.-K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *Image Processing, IEEE Transactions on*, 9(2):287–290, 2000.

[236] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, 2005.

Abstract — There is a growing interest in the development of adapted processing tools for multimodal images (several images acquired over the same scene with different characteristics). Allowing a more complete description of the scene, multimodal images are of interest in various image processing fields, but their optimal handling and exploitation raise several issues. This thesis extends hierarchical representations, a powerful tool for classical image analysis and processing, to multimodal images in order to better exploit the additional information brought by the multimodality and improve classical image processing techniques. This thesis focuses on three different multimodalities frequently encountered in the remote sensing field. We first investigate the spectral-spatial information of hyperspectral images. Based on an adapted construction and processing of the hierarchical representation, we derive a segmentation which is optimal with respect to the spectral unmixing operation. We then focus on the temporal multimodality and sequences of hyperspectral images. Using the hierarchical representation of the frames in the sequence, we propose a new method to achieve object tracking and apply it to chemical gas plume tracking in thermal infrared hyperspectral video sequences. Finally, we study the sensorial multimodality, in which images are acquired with different sensors. Relying on the concept of braids of partitions, we propose a novel methodology of image segmentation, based on an energetic minimization framework.

Keywords: Multimodality, hierarchical representation, image segmentation, energy minimization, remote sensing.

Résumé — Il y a un intérêt grandissant pour le développement d'outils de traitements adaptés aux images multimodales (plusieurs images de la même scène acquises avec différentes caractéristiques). Permettant une représentation plus complète de la scène, ces images multimodales ont de l'intérêt dans plusieurs domaines du traitement d'images, mais les exploiter et les manipuler de manière optimale soulève plusieurs questions. Cette thèse étend les représentations hiérarchiques, outil puissant pour le traitement et l'analyse d'images classiques, aux images multimodales afin de mieux exploiter l'information additionnelle apportée par la multimodalité et améliorer les techniques classiques de traitement d'images. Cette thèse se concentre sur trois différentes multimodalités fréquemment rencontrées dans le domaine de la télédétection. Nous examinons premièrement l'information spectrale-spatiale des images hyperspectrales. Une construction et un traitement adaptés de la représentation hiérarchique nous permettent de produire une carte de segmentation de l'image optimale vis-à-vis de l'opération de démélange spectrale. Nous nous concentrons ensuite sur la multimodalité temporelle, traitant des séquences d'images hyperspectrales. En utilisant les représentations hiérarchiques des différentes images de la séquence, nous proposons une nouvelle méthode pour effectuer du suivi d'objet et l'appliquons au suivi de nuages de gaz chimique dans des séquences d'images hyperspectrales dans le domaine thermique infrarouge. Finalement, nous étudions la multimodalité sensorielle, c'est-à-dire les images acquises par différents capteurs. Nous appuyant sur le concept des tresses de partitions, nous proposons une nouvelle méthodologie de segmentation se basant sur un cadre de minimisation d'énergie.

Mots clés : Multimodalité, représentation hiérarchique, segmentation d'image, minimisation d'énergie, télédétection.